Trade-offs in operating room planning for electives and emergencies: A review

Carla Van Riet*, Erik Demeulemeester

KU Leuven, Faculty of Business and Economics

Department of Decision Sciences and Information Management

Naamsestraat 69, B-3000 Leuven Belgium

Carla.VanRiet@kuleuven.be

Erik.Demeulemeester@kuleuven.be

Tel: +32-16-37.90.72

Tel: +32-16-32.69.72

*Corresponding author

Trade-offs in operating room planning for electives and emergencies: A review

The planning of the operating rooms (ORs) is a difficult process due to the different stakeholders involved. The real complexity, however, results from various sources of variability. This variability cannot be ignored since it greatly influences the trade-offs between the hospital costs and the patient waiting times. As a result, a need for policies guiding the OR manager in handling the trade-offs arises. Therefore, researchers have investigated different possibilities to incorporate non-elective patients in the schedule with the goal of maximizing both patient- and hospital-related measures. This paper reviews the literature on OR planning where both elective and non-elective patient categories are involved. It shows the various policies, the differences and similarities in the research settings and the resulting outcomes, whether they are beneficial or not. We find that the dedicated and the flexible policy are mostly pursued, but the setting and the assumptions of the reviewed papers vary widely. Decisions on both operational policies as well as on capacity are required to assure timely access and efficiency, which are the two main drivers for the problem at hand. Furthermore, the policy choice impacts the number of schedule disruptions and the OR utilization. However, results on the overtime and the patient waiting time are partly contradicting. The review shows that some policies have already received considerable attention, but the question of which policies are most appropriate is not yet fully answered. Neither has the full spectrum of policies been explored. The paper also addresses the remaining challenges for research in this field.

1. Introduction: Sources of variability

Ideally, the healthcare sector would be able to deliver the highest quality of care at the lowest cost by providing the right resources at the right time to the right patient. Unfortunately, the life of healthcare providers (and patients) is made difficult due to all kinds of variability. Examples of events inducing variability in the complete surgical process include:

- Late arrivals of patients or no-shows
- Late arrivals of medical records
- Late or early arrival of medical staff
- Delay in support services
- Inaccurate reservation of resources
- Setup, clean up or change over time variability
- Illness of patient or medical staff
- Acute onset of abnormal medical conditions (e.g., infections)
- Surgery duration variability

- Duration variability of all upstream and downstream activities (length of stay)
- Arrival of emergency patients

Many of these aspects determine whether or not the operating rooms (ORs) will run out of time or whether patients need to be cancelled. Moreover, the OR schedule influences the workload of several other departments in the hospital, such as the intensive care units (ICU), the wards, the laboratories and the Emergency Department (ED). For instance, the daily variability in the elective OR caseload is the main cause of ED diversion to other hospitals [56] or of the variability in the downstream resources [31]. Finally, emergencies are another cause of variability and need to be taken into account in order to guarantee sufficient capacity.

Litvak et al. [57] introduced the terms 'natural' variability and 'artificial' variability, which were later adopted by other authors. The former consists of variability due to the different types of diseases, each with a varying degree of illness (clinical variability), due to the unpredictable arrival of patients (flow variability) and due to the differences in the professional abilities (professional variability). This natural variability drives up the cost of care, can hardly be avoided and thus must be optimally managed. The 'artificial' variability is both non-random and non-predictable [56]. Here, many causes can be possible including patient preferences or practices of the provider. One example is the day-to-day variation in the elective scheduled caseload, which is introduced into the system by the scheduling process. This variation covers the largest part of the occupancy variation from the OR [17,57]. Artificial variability disorganizes the system since an efficient organization, where supply is nicely matched with demand, is made impossible. Haraden and Resar [31] even report that the effect of artificial variability caused by personal preferences and beliefs of the surgeons far exceeds the natural variability.

From the stochastic aspects listed before, the literature focuses mainly on the three last ones: surgery duration uncertainty, uncertainty in the length of stay (LOS) (or bed availability) and arrival of emergency patients.

First, surgery duration variability can be countered by having good estimates (e.g., [105]) or by for instance planning the expected duration increased by an amount of slack, in order to avoid overtime with a certain probability (e.g., [30]). Huschka et al. [34] show for an outpatient setting that this considerably affects patient waiting time without greatly affecting patient throughput or OR utilization. Secondly, the LOS variability can be partly reduced by having a master surgery schedule (MSS) that takes the LOS into account (e.g., [8]). As this paper focuses on the OR, the LOS will be

further excluded from the analysis of performance measures. Thirdly, the arrival variability can be tackled in different ways, which is the main focus of this review. It is important that this arrival variability both holds for electives as well as non-electives.

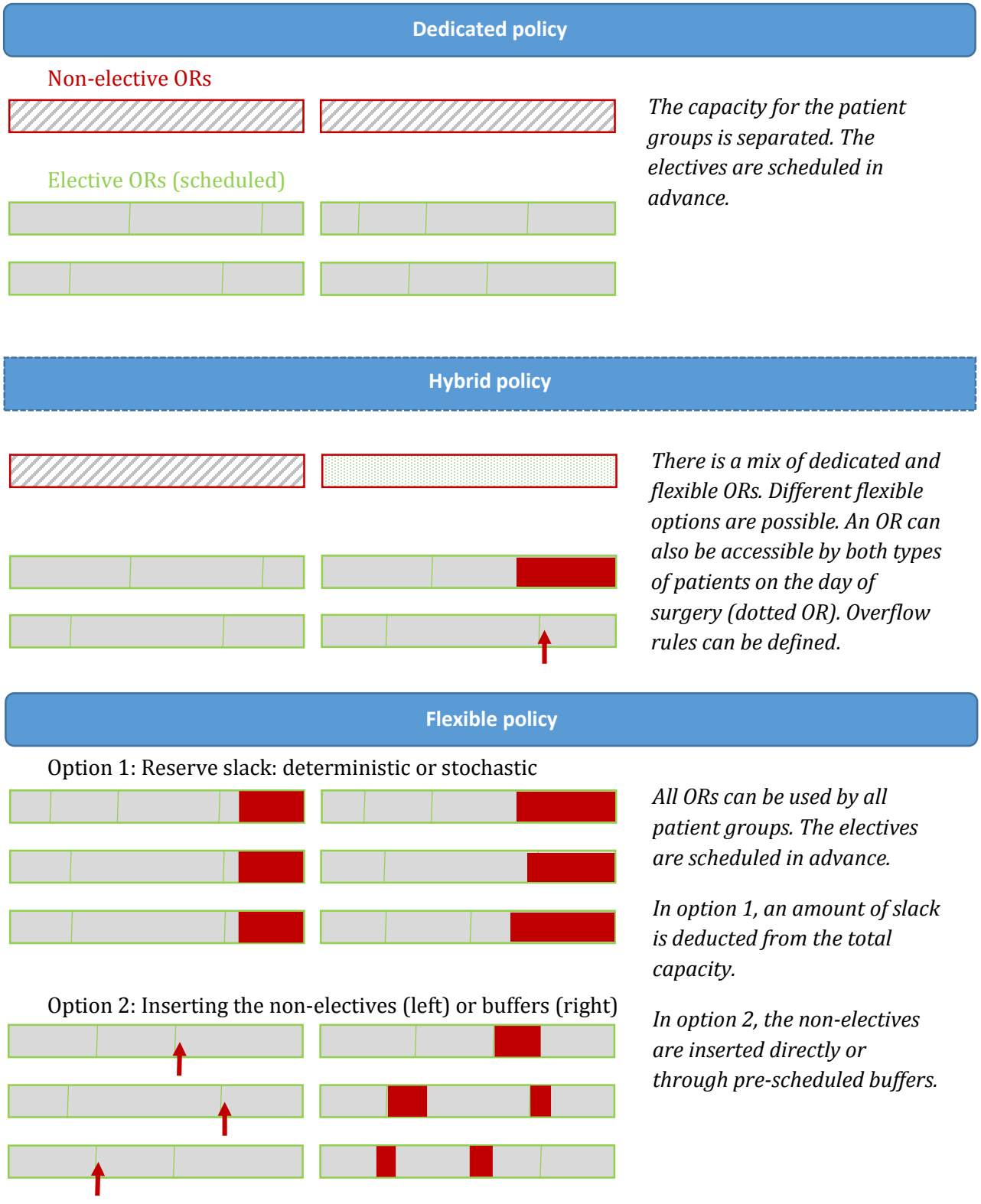
Clearly, these uncertain processes cause a need for policies to guide the decisions of the OR managers on how to manage the planning of the ORs. Unfortunately, the literature on how to include non-elective patients is scarce. Cardoen et al. [14] confirm that only limited research is being done on non-elective patient 'scheduling'. Since non-electives are intrinsically difficult to plan, most literature on operating room planning only reports scheduling practices for electives [28]. Moreover, only 29% of he reviewed papers by Guerriero and Guido [28] consider stochastic aspects. Most of these papers use tailored heuristics to overcome the computational challenges (e.g., column generation-based heuristic). Despite the research on this topic, the need for better access times for emergencies remains pressing today [67].

2. Tackling the trade-offs

The difficulty in scheduling patients in general is the trade-off between cost and efficiency on the one hand and the quality of medical care and patient's preferences on the other hand. The unpredictable nature of emergencies and the fact that they should be served on short notice creates an extra trade-off between allocating operating theater resources to non-elective patients or to elective patients. More specifically, since electives can be scheduled in advance, hospitals pursue a high efficiency level reflected in high utilization rates, acceptable patient waiting times and short turnover times. However, for emergencies, responsiveness or quick access is required. Ferrand et al. [25] discuss this trade-off in healthcare and other domains.

In order to deal with this well-known trade-off, three policies for handling emergencies are pursued: the flexible, the dedicated and the hybrid policy. In the flexible policy, there is no separate OR reserved for non-electives and several rules and scheduling strategies are used in order to manage the access for the two patient categories. Both advanced scheduling strategies as well as operational strategies for on the day of surgery must be defined. In the dedicated policy, one or more ORs are dedicated to a specific patient type in order to separate the flows of the patient categories. The hybrid policy is a combination of both in which for instance some capacity is reserved for non-electives, but other ORs are also accessible by non-electives. The different policies, illustrated in Figure 1, are further discussed in more detail.

Figure 1: Illustration of the three different policies for handling non-electives and electives



3. Organization of the review

We searched the database Web of Knowledge for relevant manuscripts, written in English and appearing between 1990 and2014 in the areas of operations research management science and health care sciences services. Search phrases included: emergent surgery planning/scheduling, emergency theater, semi-urgent surgery planning, urgent surgery planning/scheduling, non-elective patient scheduling, emergency operating room, dedicated operating room capacity and operating room capacity emergency. Furthermore, other relevant papers were selected based on reference list checking and the criteria explained below.

This review focuses on the OR literature that directly impacts or explicitly considers non-elective surgeries. More specifically, this means that the non-elective patient category should be taken into account in the operational or tactical decision making. It includes both research on tactical allocations (capacity) as well as on operational patient scheduling. Mainly papers that use (technical) operations research techniques (mathematical modelling, simulation) are discussed. Papers reporting on data-analysis are included if the focus is on a comparison between (the implementation of) different policies. Other managerial papers are not classified, but are mentioned when they provide specific insights. Other techniques such as workflow management and business process re-engineering (e.g., [6]) are not included. Papers that deal exclusively with non-electives in the up- and downstream resources such as the ED and the ICU (e.g., [58]) are not included in the classifications. Finally, disaster management is also considered to fall outside the scope of this review.

Only a few papers make an enhanced comparison between a flexible and a dedicated policy (e.g., [23,51,89,103]). In section 4, they are classified under both policies, while in sections 5 and 6 they are classified only in the section where they are assessed to be most relevant. When a paper discusses more than one policy, it is classified under the policy that receives most attention in the paper.

In the patient scheduling literature, there is a lack of consistent designation of patient categories. The following terms are all used to describe patients who cannot be scheduled well in advance: emergent, urgent, add-on, work-in and semi-urgent. We will further use the term non-electives throughout this review to address this patient group.

The rest of the review is structured as follows. Section 4 provides an overview of the characteristics of the building blocks that are used in reviewed papers. Section 5 discusses the literature on the flexible policy and sections 6 and 7 treat its dedicated and hybrid counterpart respectively. Challenges for researchers and conclusions are addressed in the final section.

4. Characteristics of the building blocks

When comparing papers, the researched setting and the corresponding assumptions are important. This section describes these building blocks of the literature on non-electives in the ORs. The first two subsections discuss the policies in relation to respectively the scope of the research combined with the size of the case hospital, and the time window of the dataset combined with the decision level. The third subsection looks at the modeling assumptions. Since categorization and prioritization are important aspects for non-elective 'scheduling', the fifth subsection clarifies them, followed by a review on the ratio of non-electives to elective patients. Finally, the type of analysis and the applied solution techniques are summarized.

4.1. Scope and OR size

Most of the reviewed papers look into a problem of a specific case hospital. Therefore, it is important to look at the scope of the research, in the sense of the departments that are involved, and the applied policy simultaneously, as shown in Table 1. Furthermore, the size of the OR complex of the hospital greatly influences the results as confirmed by van Essen et al. [92] and is therefore shown in Table 1. The number of ORs has an impact on scalability, pooling options etc. Note that although the number of beds is regularly used to depict the size of the hospital, Table 1 shows the number of ORs since this is more relevant in our context.

Table 1: The applied policy in relation to the scope of the research and the OR size

	Scope (case hospital)				
Policy	**All departments**			**Specialized**	**No clear/real data**
OR size	*<10 ORs*	*10-15 ORs*	*>15 ORs*		
Dedicated		[55,69,103]	[23,32,75,79]	[9,71,74]	[7,18,59]
Flexible					
Option 1	[22,73]	[64,94,103]	[23,30,95]	[1,106]	[26,46,47,48,49]
Option 2	[19,72][a]		[19,92]		[84]

[a] Two hospitals are researched in [19].

More than half of the papers, classified in Table 1, research a setting with multiple departments. This setting or scope is different from the one focusing on one or a few departments. Clearly, different departments or pathologies have different characteristics in terms of arrival patterns, duration patterns (both mean and variance) and allocated capacity and staff. Isolating a pathology can therefore greatly influence the results and even more importantly, limit the results to the characteristics of this pathology. Additionally, 27% of the papers fail to report clearly on this aspect.

With regards to the policies, both the dedicated and the flexible policy have received attention in the literature, with slightly more focus on the flexible policy.

Looking at more detail of the flexible policy, Table 1 shows that the second option of the flexible policy has received less attention than the second one, since 75% of the papers on the flexible policy discuss the first option. A similar analysis for the OR size shows that although all three categories in Table 1 are represented in the literature, the majority of the papers considering all departments, research an OR complex of reasonably large size (> 10 ORs).

The interaction between the different classification factors shows that overall the dedicated policy in a setting of specialized departments or a large OR complex has received the most attention. In the majority of the cases looking at specialized departments, the orthopedic department is discussed. Since the specialized departments might already be selected because of their favorable characteristics, results have to be interpreted with caution. Moreover, a lack of attention for flexible option 2 is apparent in most settings. Additionally, the small OR setting has gained little attention for both policies. This might result from the idea that a small OR complex provides fewer possibilities for resource dedication compared to a hospital with more ORs.

In summary, the dispersed classification of the papers in Table 1 partly provides a reason for the contradicting results on the policies in the literature, especially when it comes to the question of whether or not to dedicate ORs. Moreover, the second option of the flexible policy provides an opportunity for future research, since it has received limited attention so far. Especially research on inserting variable-sized breaks at several spots in the schedule is scarce. Moreover, only one paper [19] incorporates real data of hospitals with different OR sizes.

4.2. Data time window and decision level

As mentioned before, the two main policies for handling non-electives are studied at both the operational decision level as well as at the capacity decision level, as reflected in Table 2. Once the decision level is known, the presented data must be examined carefully. Unfortunately, they are often not fully disclosed, which makes it hard to check the quality of the dataset. Nevertheless, the time-window covered by the data can be an aspect to keep in mind when interpreting the results and is therefore shown in Table 2.

Table 2: Time window of data and decision level per policy

	Operational		**Capacity**	
Time window of data	*< year*	*>= year*	*< year*	*>= year*
Dedicated	-	[71]	[23,32,59]	[7,9,55,69,71,74,75,79,103]
Flexible	[19,72]	[1,30,92,94]	[23,106]	[1,94,103]

Note: [18,22,64] consider an unclear time window.

At the capacity level, the focus is on having fast access to care for the non-elective patients while minimizing the disruptions for elective patients. Operational studies focus more on the best way to schedule both patient groups. Some papers discuss both levels and focus on the operational scheduling policy on the one hand and the number of dedicated ORs (DORs) or capacity on the other hand. Overall, most research focuses on the tactical decision level (e.g., how many ORs need to be dedicated). Operational policies to manage the non-electives in the DORs are lacking and form an area for future research.

Table 2 shows that the majority of the classified papers (71%) work with a dataset covering at least one year. Of these papers, the majority works with data of exactly one year.

4.3. Modelling assumptions: Surgery duration and patient arrivals

Variability in durations and arrivals is causing scheduling difficulties. The corresponding assumptions determine the range of analytical possibilities and how well the model resembles reality. Table 3 provides an overview of the assumptions that are used to model the surgery durations and the patient arrivals for both elective and non-elective patients. Authors that are not mentioned in the table did not include information about the distributions. The table also considers the assumptions made in testing phases.

With regards to arrival times, two approaches can be discerned. Firstly, the expected number of arrivals per time unit, also called the arrival rate, can be used to model the patient arrival process. This arrival process is generally modelled as a Poisson process. The Poisson distribution implies that the mean arrival rate equals the variance.

A second approach to model the arrivals uses the inter-arrival times. The inter-arrival times are especially popular in queueing and simulation models. Following the results from the previous paragraph, the exponential distribution is often assumed. Note that for non-electives detailed arrival times are relevant, while for elective patients the arrival time can be modeled with a more aggregated

time unit. The use of the exponential distribution brings along several benefits which (partly) explains its popularity. As such, the parameter for the distribution can easily be calculated by taking the inverse of the mean. Moreover, several theoretical queuing results are available for systems with exponential inter-arrival times. Finally, sometimes fixed inter-arrival times are assumed (e.g., arrivals at the start of the day), but the type of arrival is stochastic (e.g., [22,72]).

The patient arrivals of electives in Table 3 must be interpreted with caution. In several papers on the operational scheduling, the elective schedule is assumed to be fixed (deterministic). In an outpatient setting, this means that the scheduled starting times of the patients are known, while in an inpatient setting this might be reduced to knowing the sequence of surgeries. Another option is that only the number of patients (based on historical data) is known. Therefore, the arrival section for electives in Table 3 only shows the papers that make assumptions on the inter-arrival times.

Table 3: Modelling assumptions on durations and arrivals for both electives and non-electives

	Surgery duration		**Patient arrival**	
Elective	• Lognormal	[23,24,46,71,73,81, 84,92,103]	• Poisson	[1,26,71,104]
	• Deterministic *Historical mean*	[1,72] [22,71,92]	• Deterministic	[22,23,24]
	• Empirical distribution	[64,75,76,104]		
	• Uniform	[47,48,49]		
	• Normal[a]	[26,30,64,81,94,95]		
	• Exponential	[88,89]		
Non-elective	• Lognormal	[19,23,24,46,71,81, 84,92,103]	• Poisson	[1,11,23,24,65,69,71,76 ,88,89,92,103,104,106]
	• Empirical distribution	[11,65,75,76,104]	• Other	[22,32,72]
	• Erlang	[69]		
	• Normal[a]	[26,47,81,94,95]		
	• Exponential	[47,48,49,88,89]		

Note. Papers based on data-analysis (see Table 7 in section 4.6) report the realized durations and arrivals and are therefore not included while papers on the hybrid policy are. Two entries in the table can relate to modelling both the expected and realized durations or to different assumptions for the proposed model and the (stochastic) testing phase.
[a] Mostly the assumption relates to the sum of the durations within an OR day or the sum of the non-electives. Similarly, [73] model the number of OR hours for non-electives as a Poisson distribution.

With regards to the surgery duration, Table 3 suggests that the lognormal distribution often results in the best fit for the hospital data. The paper by Strum et al. [83] is often cited to confirm this assumption. Moreover, the normal distribution is for both patient categories the second most popular

one. Finally, deterministic surgery times for electives (including the historical mean) are regularly assumed.

The estimation of the surgery duration remains a topic that requires attention in practice. Lebowitz [52] even reports inaccurate estimations as the most important cause of the lack of OR punctuality. On the contrary, Dexter et al. [17] show that optimally choosing the operating day for the elective patients is more important in order to best fill the allocated hours compared to eliminating the errors in the estimation of the surgery duration. A detailed analysis of surgery duration estimation is out of the scope of this paper, but developing forecasting tools remains an area for future research.

4.4. Categorization and prioritization

If we refer to different patient categories, the question of how to categorize and prioritize them raises naturally. Litvak et al. [57] argue to separate the different sources of variability by classifying the patients into homogeneous subgroups. Both the categorization of patients and the prioritization between patients play an important role.

Finding a good basis for categorization is crucial since a trade-off exists between separation and flow variability and one cannot endlessly divide the patients into smaller subgroups. Examples for all patient groups are disease type (e.g., Cardiology, Orthopedic), attending surgeon, disease severity (e.g., complex or simple) or resource type (outpatient or inpatient). Moreover, duration categories can be used to separate patient groups (e.g., [54]). Sandbaek et al. [77] show that introducing a new classification system resulted in a system where 90% of the patients could be planned ahead.

Although a classification system can be important to decrease the waiting time for non-electives, these patients are often categorized under a general term like ‘emergent’ or ‘urgent’. Especially for the add-on cases a common definition is lacking, since this term is used both as a collective term (e.g., [72]) as well as to categorize a specific type of patient (e.g., [75]) or for everyone that is added after a threshold hour for surgery on the next day (e.g., [17,32]). The vague and inconsistent categorization makes it more difficult to compare or benchmark papers in this field.

Often categorization is based on medical urgency (elective or non-elective patient). Many papers (and thus case hospitals) use the medical priority as main categorization basis. A recent example of categorization based on medical priority is one where patients are assigned a time interval in which their surgery is medically advised, called the due time [97]. This categorization recognizes that

scheduling lower-priority patients later in the time horizon provides additional flexibility to schedule higher-priority patients [68].

A wide range of medical categorizations is used in the literature. Table 4 provides the different categories and the accompanying service target. Unfortunately, the same category is sometimes used for denoting patients with varying service targets and often no (interval) target time is reported. The intervals can positively influence the number of non-electives treated within their deadline as shown by van Oostrum et al. [96] where safety intervals are introduced during the night shift. Moreover, the target time can provide hospitals a way to measure the category-specific waiting time performance. The due time concept also turns up in Canadian, British and Italian research (e.g., [90] use five categories for elective surgery).

Table 4: Categorization of the non-electives

Category	**Target**	**Reference**
Trauma	Now	[11]
Emergent	<30min	[23,24,75]
	<1h	[7,22,69]
	<2h	[69,72,75]
	<6h	[9]
	<24h	[26,71,73,88,89,104]
Urgent	<4h	[69,75]
	<24h	[7,9,72]
Semi-urgent	<8h	[69,75]
	<1/2w	[106]
Add-on	<24h	[9,72,75]
Add(-on) elective	No target: fill up free capacity	[18,72]
Non-urgent	<24h	[69]
Work-in	[24h–1w]	[75,79]
Other		
Priority levels (P)	P1-P3 (emergent): <1h, <4h, <12h	[32]
	P1-P3: <8h, <8-24h, <24-48h	[55]
	P1-P3: <6h, <24h, <78h	[77]
	P1-P5 (emergent): <45min, 2h, 4h, 8h, 24h	[79]

Note. w= weeks; h = hours; min = minutes; d = days.

Conversely, questions can be raised about the quality of the common medical categorization, as reflected in Table 4. In practice, most of the non-electives can actually be treated within 24 hours. Typically only 15% of the non-elective patients must be served within six hours of admission [50]. Similarly, Heng and Wright [32] argue that only 9% of the patients are of the highest urgency category

(i.e., requiring surgery within one hour) while 63% can get surgery within twelve hours. Even in the orthopedic trauma center researched by Bower and Mould [11], the real trauma patients are only 25% of the total non-elective patients. This mismatch is reflected in results that are measuring whether or not the patients are served within their due time. For instance, Leppäniemi and Jousela [55] show that the highest priority non-electives are usually (in more than 80% of the cases) served within their target time, while this number decreases for the medium and lower priority cases. Nevertheless, Samudra [76] reports opposite trends, where respectively 74%, 78% and 83% of the three highest urgency categories are served within their deadline.

The categories can be used to set priorities between patients. To clarify this, it is important to note the distinction between urgency and priority. Urgency is primarily based on medical criteria at the time of arrival of the patient to the hospital. Priority, however, refers to the relative position of the patient with respect to other patients on the waiting list and is often used to develop an admission rule. Sporadically, multiple non-elective patients arrive approximately at the same time so that next to categorization, priorities within a patient category need to be set to organize the surgical process.

In general, defining the sequence of patients (or priority) can happen according to several criteria: the position in the waiting list (leading to e.g., first-come, first-served (FCFS)), patient specific characteristics (e.g., type, duration, urgency) or the contribution to an objective function or to a combination of factors (e.g., priority scores). Table 5 shows that position is used in several models and simulations, even though the survey of Cardoen et al. [15] report that about 68% of the respondents indicate that the arrival sequence is a less important factor in determining the priorities. Although several authors [9,32,69,106] explicitly mention that the non-electives are served according to the applied classification system (i.e., there is a predetermined priority between the different categories), most papers do not mention the prioritization system. Dexter et al. [18] propose to prioritize urgent surgeries in increasing order of expected surgery durations if the scheduling objective is to minimize the average length of time each patient waits and show also the results for FCFS and medical priority prioritization.

Table 5: Prioritization rules for non-electives

Prioritization basis	**Reference**
Position in waiting list	[1,18,23,24,92,103]
Patient characteristic (type, duration, urgency)	[18,32,69,84,106]
Score (combination of criteria)	[71]

Priority scores thus try to take multiple factors into account. They are based on a wide variety of aspects often including medical urgency, professional priorities, resource use and time spent on the waiting list. Mullen [66] provides a comprehensive review of the priority scores developed over time including additive and multiplicative forms. The objective of using priority scores ranges from determining whether a patient is delayed or denied to defining the patient's urgency and importance. In addition, MacCormick et al. [60] published a review on prioritization systems of elective patients in which they discuss the different factors and their weighing. They show that only 13 out of the 50 reviewed studies include recommended waiting times together with a prioritization system.

Testi et al. [90] compare a scoring algorithm, that determines a relative priority for each patient in the waiting list, to a clinical assessment with a recommended maximum waiting time. They propose the priority score in Formula (1) where c is the urgency status that is calculated as the weighted sum of the numerical values of three clinical criteria. They use the need-adjusted-waiting days as a performance measure to include both urgency and priority and conclude that both methods should be used simultaneously. Indeed, Sobolev et al. [80] show that less urgent patients might have a higher probability of admission in relation to their waiting time compared to more urgent cases if a classification system based on urgency is used. This is possibly caused by hospital-related or patient-related delays.

$$Priority\ score = c * Waiting\ Time \tag{1}$$

Finally, to further schedule the OR (after prioritizing the waiting list), bin-packing algorithms such as best-fit (e.g.,[19]) can be used in order to satisfy certain objectives. Note that FCFS prioritization might in reality result in an order where a patient who is arriving first, gets the first *feasible* slot (i.e., a slot that has enough free capacity to serve the patient) and might thus be served later than a patient arriving later, but having characteristics that allows him/her to be scheduled earlier (e.g., shorter duration). Prioritization can be useful for both the patient-to-day assignment (primarily in hospitals that schedule based on a waiting list) as well as well as for the sequence within an OR-day. For non-electives, the latter is most relevant.

4.5. Patient mix

Non-elective patients follow a specific flow in a hospital. They usually enter the OR complex from a variety of locations (e.g., the ED, the inpatient wards, the ICU or an external hospital). Once the

surgery request arrives, an OR and a surgical team are booked and the elective schedule might need to be adapted. After surgery, non-electives follow a clinical pathway similar to the one of electives

In order to 'schedule' non-electives and trade-off capacity between non-electives and electives, it is relevant to know the patient mix or the probability that a non-elective patient will arrive. Since non-electives might come from the ED, it is interesting to first look at the admission rates for ED patients to predict the flow of incoming non-elective patients. Peck et al. [70] studied four hospitals and the admission rate varied from 26% to 32% of the approximate monthly ED patient volume. However, no daily or hourly rates are provided. Leppäniemi and Jousela [55] report a 50% admission rate.

Another informative number is the percentage of patients that receive surgery and enter the OR as a non-elective. Table 6 shows the ratio of non-elective patients to the total patients and displays a wide variety in the patient mix of different case hospitals. Unfortunately, as shown in Table 6, only a few papers discuss this ratio, despite its importance in capacity decisions.

Table 6: Number of non-elective patients as a percentage of the total admitted patients

Category	**%**	**Reference**
Non-elective	10-15%	[75]
	14%	[23,24,103]
	17%	[89]
	22%	[76]
	25%	[1,59]
	33%	[77]
	50%	[55]
Semi-urgent	40%	[106]
Add-on	20%	[32]

With regards to add-on cases, Dexter et al. [20], defining an add-on case as a patient that is scheduled after 7 PM for the next day, shows that 24% of the slots (i.e., a combination of an OR and a date) contain add-on cases for their case hospital. In addition, at least half of the ORs have the last case scheduled or changed within two days of surgery. Moreover, Dexter et al. [17] also show that the percentage of add-on cases can differ greatly between hospitals and especially between an outpatient (6%) and a university hospital (16%) setting.

4.6. Type of analysis

The overview of applied solution methods and the corresponding type of analysis are summarized in Table 7. In the more medically oriented journals, data-analysis and simulation are often used to tackle OR problems of a specific case hospital. Often the pre- and post-implementation situation of a specific OR policy in the case hospital is examined. In the papers focusing on operations research methods, simulation, mathematical programming, heuristics and analytical analysis are mostly used. The following paragraphs discuss the popular operations research methods in more detail.

Table 7: Type of analysis and solution methods

Type of analysis	
Scenario analysis/modelling	[1,11,23,24,30,32,51,62,64,65,69,71,73,75,76,79,81,82,88,89,90,92,94,95,103,104]
Retrospective study	[7,9,32,55,59,74,75,77,79]
Optimization	
Exact	[1,26,43,48,51,69,71,72,73,84,92,104]
Heuristic	[1,18,19,22,30,46,47,49,51,64,65,81,92,95]
Complexity	[30,46,47,48,49,51,92]
Solution method	
Simulation	
Discrete event simulation	[1,19,23,24,51,64,69,71,75,76,92,103,104]
Monte Carlo simulation	[11,46,47,48,64]
Data analysis/observation	[7,9,32,55,59,74,75,77]
Analytical analysis	
Queuing/Markov theory	[32,69,88,89,106]
Other	[47,48,69,94]
Constructive heuristic	[18,19,30,46,47,49,51,92,95]
Improvement heuristic	
Meta-heuristic	[22,30,47,51,92]
Other	[30,46,47,49,51,81]
Mathematical programming	
Goal programming	[1]
Column generation	[46,49]
Mixed integer programming	[22,46,47,48,64,71,72,73,92,104]
Linear programming	[18]
Branch and bound	[84]
Dynamic programming/MDP[a]	[26,49,65,106]
Unclear	[79]

Note. The classification schema is based on the one of Cardoen et al. [14]. Since [92] and [103] are follow-up articles on a working paper [51], the last one will not be classified separately in the remainder of the paper.
[a] MDP = Markov Decision Process

Discrete event simulation (DES) is a common approach for studying complex environments and for taking uncertainty into account. Not surprisingly, it has been increasingly popular as a tool to

tackle problems in the operating theatre, which is reflected in the increasing number of scientific contributions using this method. As shown in Table 7, simulation is also commonly used to model the OR scheduling problem. It is used for modeling the (non-)elective patient flow, the resource allocation, patient scheduling and capacity decisions. This result fits in the findings of Brailsford et al. [13], who show that simulation is prominent in planning and resource utilization problems.

Several reviews on simulation modelling in general (also including Monte Carlo simulation, system dynamics etc.) in healthcare are published (e.g., [13,36,37,63]). Moreover, Augusto et al. [4] provide a framework or meta-model for simulations of healthcare systems and Jahangirian et al. [35] discuss the lessons learned from the commerce and defense sector that might be applicable for simulation in healthcare.

Whether or not to use simulation depends on the research objectives. The goal to incorporate the stochastic elements of the ORs as precisely as possible often motivates the use of simulation models. Simulation can provide answers to particular questions of the case hospital and is suited for scenario analysis. On the contrary, simulation models often do not provide optimal solutions and might be computationally expensive.

Moreover, Brailsford [12] discusses the barriers to implementation of the simulation models in healthcare and Jahangirian et al. [35] show that the sector is lagging behind on that aspect compared to the other two discussed sectors. Both papers mention the resistance to (organizational) change and the data capture as causes for failed projects.

Mathematical programming and mixed integer programming (MIP) in specific is a popular optimization technique. It is mainly used in the reviewed papers to create optimal operational surgical schedules. As a main benefit, it can yield solutions that are optimal, given the input. However, for some datasets, the current techniques are not able to solve the problem in a reasonable amount of time (where reasonable must be judged according to the problem setting).

Heuristics are often seen as the tool to overcome the intractability of a mathematical program, without the need for an expensive simulation model and are therefore also popular for constructing OR schedules. Since the patient-to-OR-day problem can be formulated as a bin-packing problem, many heuristics are proposed to solve this well-known problem. Examples are the best-fit and worst-fit heuristic. Heuristics are easy to implement at the cost of loss of optimality. Moreover, it is not always easy to find bounds for approximation algorithms.

Queueing analysis has been applied to various healthcare problems, especially the appointment scheduling problem. There are several good reviews of queueing theory applied in healthcare [27,29]. Queueing analysis is not only suitable for assessing the operational performance of a system, but is also used for determining capacity requirements, such as determining the required number of ORs for certain patient categories. Both applications are found in the non-elective literature. Compared to simulation modeling, this method needs less data and often might provide a closed form formula to calculate performance metrics. McManus et al. [62] show for instance that queueing theory is suited for modeling critical care resources. Also trade-offs between OR utilization and overtime or cancellations are modelled using queuing theory (e.g., [106]). However, often restricting assumptions on the distributions of the arrivals and the service times are necessary and steady-state models (stationary service) cannot be applied to all OR settings.

More specifically, Markov models (both Markov chains and Markov decision problems (MDP)) can be used to model or optimize healthcare problems. A problem is Markovian if the future state depends only on the current state and not on states preceding the current state (i.e., the feature of memorylessness). Markov chains are used to model patient flow, can be used as statistical models of a system and are popular in for instance the ED literature. Tancrez et al. [89] use Markov theory to guide decisions on the size and the allocation of capacity to (non)-electives. An MDP is essentially a sequential decision model and adds decisions and rewards to the Markov chains. Zonderland et al. [106] use an MDP to determine at the beginning of a particular week how many slots should be planned in that week for the so-called two-week semi-urgencies, which are urgencies that must be served within two weeks.

5. Flexible policy

In a flexible policy there is one pool of ORs in which all patients, both non-elective as well as elective, are operated.

As explained earlier in Figure 1, a flexible policy contains two options. A first option, used by many hospitals, consists of filling the OR for a given fraction of the full capacity (e.g., 85%) to leave some safety margin or slack for unexpected events (e.g., arrival of non-electives). Clearly, the safety margin decreases the OR capacity which is assigned to electives.

The slack is mostly planned ‘virtually’ at the end of the day. It can be either a stochastic or a deterministic amount. In the deterministic case, the slack is for instance based on the expected

number of non-elective arrivals. In the papers advocating the stochastic approach, the non-electives are usually handled as a separate, aggregated category without specifying the different urgency levels. This approach is mostly preferred in the papers using a mathematical optimization model.

The other option is to schedule specific moments throughout the day at which non-electives can enter the elective schedule. These moments can be either break-in-moments (BIMs) [92] or time intervals or breaks. The idea of BIMs is to minimize the time that an arriving non-elective patient has to wait before receiving surgery. Note that no capacity is left free in this case, only a possibility for entering the schedule is created. Alternatively, inserting breaks does leave capacity free at predefined spots in the schedule.

According to a survey of Cardoen et al [15], most hospitals (85%) in Flanders adopt the flexible strategy and plan non-electives in the first OR that becomes available. With regards to urgencies, the majority of the respondents incorporated these patients in the regular program of the appropriate discipline during the day (54%). Another 30% uses the practice of operating the urgent surgeries at the end of the day program and the remaining 16% combines both practices.

5.1. Results

The advocates of the flexible policy mainly report on five performance measures as reflected in Table 8 and most papers consider more than one. The direction of the performance measures when moving towards a more flexible policy, shown in Table 8, is based on those studies comparing different policies or on explicit notes on the direction in other papers. The same reasoning holds for the dedicated policy in section 6.

The results on staff overtime are contradictory. Some authors report a possible increase due to the increased variability [23] in the schedule while others note a reduction in the overtime [103]. Wullink et al. [103] show that the average overtime per day decreases, but the average number of ORs with overtime per day is (slightly) higher in the flexible strategy. On the contrary, Ferrand et al. [23] indicate that although the average overtime increases, the number of patients served in overtime remains the same.

The results for waiting time are different for both patient categories. The elective patient waiting time increases according to most authors, because the elective schedule is disturbed by the incoming non-electives. According to the papers advocating a flexible approach, pursuing a more flexible strategy results in a lower waiting time for non-electives. The reasoning behind this is that the non-

elective patients have more possibilities to enter the ORs since they can access all ORs. However, proponents of the dedicated strategy argue that the opposite is true, as discussed in section 6.

With regards to waiting time, it is important to keep the different types of waiting time in mind. The time between the decision for surgery (often made after a consultation) and the actual day of surgery is called indirect waiting time and is only relevant for electives. Direct waiting time refers to the time a patient has to wait longer than its scheduled time (or arrival time for emergencies). While the former metric is important in evaluating the performance of the patient-to-date scheduling algorithm, the latter is more relevant for sequencing decisions and the access time for non-electives.

Furthermore, the flexible policy results in an improved overall OR utilization and in an increased number of cancellations of electives. The flexible policy avoids that an OR remains idle because there are no patients requesting dedicated OR time. Surprisingly, the cost of cancellation is overlooked in many papers which might give a biased view on the results of the flexible policy. After all, cancellations are the other side of the trade-off between utilization and overtime. Moreover, since electives are usually admitted to a ward before having surgery, the non-elective arrivals mainly cause inconvenience for the electives, rather than disturbing the processes in the OR [103].

Table 8: Performance measures in a flexible policy

Performance measure	Increase/decrease	Reference
Waiting time electives	↑	[1]
Waiting time non-electives	↓	[92,103]
Staff overtime	No consensus	[1,22,26,30,46,47,48,49,64,65,73,94,103]
Utilization of OR (overall)	↑	[19,30,49,94,95,103,106]
Cancellations/rescheduling	↑	[22,26,64,84,106]

In addition, other performance measures show up in the reviewed papers. As such, some authors consider the waiting time for both patient groups as one performance measure (e.g., [1,65,73]). Also the leveling of resources or patient volume (e.g., [1,92]), the number of deferrals (e.g., [1,73]), the makespan of the OR session (e.g., [72]), the bulking cost (e.g., [65]), OR undertime (e.g., [106]), priority scores (e.g., [64]) or general elective related cost (e.g., [46,48,49]) are considered. Adan et al. [1], for instance, show that more flexibility decreases the overall waiting time, but their setting focuses on flexibility with regards to scheduling electives.

Interestingly, van der Lans et al. [51] argue that a higher process variability leads to lower non-elective waiting time in BIM optimization and Ferrand et al. [23] suggest that in a highly variable system, going towards a more flexible system might benefit both patient categories.

In general, in the flexible approach of reserving slack, the main goal seems the trade-off between utilization and overtime, while for the break-in-moments non-elective waiting time is the main driver. A practical drawback of the flexible policy with reserving capacity is that a successful implementation requires everyone's collaboration since everyone has to reserve free capacity.

5.2. The handling of non-electives

Non-electives arriving during the day who could not be fitted into the schedule are operated on in overtime or during the night and evening shifts. Lovett et al. [59] report that 53% of the non-electives are served outside regular hours (i.e., after 5 PM in the case hospital).

However, non-electives should be served as close as possible to their arrival time. In order to reach this goal, decisions need to be made both in advance (the advance scheduling phase and the tactical planning) as well as on (or close to) the day of surgery. First of all, a decision on whether or not to introduce slack time in the flexible ORs needs to be made, together with the size of this slack. Secondly, scheduling algorithms for inserting the non-electives in the schedule must be determined.

Managing the slack

First, consider the option of reserving a specific amount of capacity (slack) without scheduling this capacity (option 1 in Figure 1).

Van Houdenhoven et al. [94] calculate the reserved capacity per day based on the average duration and corresponding variance for non-electives and electives and the desired risk of overtime. A norm utilization is calculated per surgical department in order to address the trade-off between OR utilization, case mix and accepted risk of overtime. The optimal slack time depends thus on the case mix and the block length. They assume that the total surgery duration is normally distributed. Wullink et al. [103] distribute the amount of slack equivalent to one OR-day evenly over the ORs. Each specialty reserves one emergency surgeon per day (e.g., on a day research or administrative tasks are planned for this surgeon). Unfortunately, they do not consider the waiting time of elective patients. Adan et al. [1] calculate the reserved slack based on the arrival rate of non-electives on a specific day and the probability that the non-elective patient arrives at daytime. They determine the capacities per specialty with a goal programming approach. They schedule a number of patients that is higher than the average in order to create slack in the operational plan, since the actual number of arriving patients is not equal to the average number of patients in the tactical plan.

The amount of slack can also be determined by queuing theory as done by Zonderland et al. [106]. The authors examine the trade-off for semi-urgent patients between societal cost (due to patient cancellation and waiting time) and the unused capacity. They determine the minimum amount of OR time to reserve based on the number of patients of each type arriving per week and the number of required slots for each type and develop a slotted queuing model in discrete time. The slots are assumed to be of equal length. They calculate the distribution of the number of slots requested at the beginning of each week. The authors argue that the number of cancellations might increase if only the average behavior of the system (minimum slack) is taken into account.

Other authors represent the slack as a stochastic variable. Lamiri et al. [46,47,48,49] provide an elective surgery plan including uncertainties in the form of random variables for surgery durations and for the required capacity used by non-elective arrivals. They mostly assume exponential durations for non-elective and uniform distributed durations for elective patients. However, in [47], they test different distributions. Also Min and Yih [64] incorporate the capacity used by non-electives as a random variable in a stochastic MIP model. Furthermore, Gerchak et al. [26] also present the amount needed for non-electives as a random variable and look for the optimal amount of capacity to reserve for non-elective patients each day. The amount of slack is thus dynamic and depends on the size of the list of patients waiting for surgery on that day. They show that this amount is a decreasing function of the number of elective patients waiting. Moreover, Rachuba and Werners [73] ensure that the stochastic amount required for non-electives is divided over a maximum number of ORs.

The possibilities for the scheduling algorithm (electives) are influenced by the amount of slack and often both are interrelated. For instance, Van Houdenhoven et al. [95] and Hans et al. [30] use the portfolio effect, which states that portfolio risk decreases with increasing diversity (i.e., no correlation between components exists), and cluster the surgical cases with similar variances. This results in less total planned slack. For instance, a slack time of 40 min would be optimal for ophthalmology while for ear, nose and throat this would be 110 min. Hans et al. [30] test a combination of constructive heuristics and local search techniques. They want to maximize the total free capacity and thus to minimize the required slack, which is dependent on the characteristics of the individual surgeries and on the amount of surgeries that are already planned in the OR. They show that regret-based random sampling is the best constructive approach. The solutions can be further improved by the random exchange method.

Based on the chosen amount of slack, a Markov decision model is used in [106] to determine at the start of each week the number of slots for the two-week semi-urgencies to plan in week one. If a

one-week semi-urgent patient arrives, first the reserved one-week semi-urgent OR time is used. Then, if this time is not sufficient, electives are cancelled and only then overtime is used. Cancelled electives become semi-urgent patients, which need to be served within the following two weeks.

Lamiri et al. focus both on the patient-to-date [46,47,48,49] and the patient-to-OR scheduling problem [47,49]. They formulate the planning problem as a stochastic mathematical program and solve it by first approximating the MIP formulation by the Monte Carlo sampling technique. Later, they extend the model by applying column generation, where each column represents a possible assignment of elective patients to an OR. Heuristics are used to derive a feasible solution to the integer problem, which is then improved by local search. They do not include waiting time explicitly, but the models penalize for delaying the scheduling of electives. As a second step in their approach, Adan et al. [1] fill up the planned slots created earlier in the operational phase with or without following several flexibility rules (e.g., allowing patients from other specialties to be booked in the OR). Afterwards, the execution of non-elective and elective surgeries happens according to the daily scheduling algorithm. They clearly show the trade-off between hospital efficiency and patient satisfaction (e.g., waiting time, cancellations) along the efficient frontier.

Opting for slack means that on the day of surgery, often decisions on how to deal with the arriving non-elective still need to be made. For instance, Adan et al. [1] apply the following daily scheduling algorithm for dealing with non-electives. Non-electives that arrive during night time are operated on during the night shift and the beginning of the day shift if necessary. If they arrive during the day, they are operated on during the day (after the elective program) and at the beginning of the night shift (if needed). For each arriving non-elective, the decision rules for cancellation (i.e., deferral) are followed. The cancellation decision is based on the estimated required resources at the beginning of the day. Electives that are 'cancelled' are postponed to the same slot in the next week. Non-electives can also be served at the end of the day, after the elective schedule is finished [78]. Wullink et al. [103] apply the first-fit algorithm for scheduling the electives, based on the historical mean duration. Ferrand et al. [24] use fixed inter-arrival times for the elective schedule.

Going one step further at the operational level leads us to online (re)scheduling where scheduling (i.e., defining the starting time) happens upon the event of an arrival of a patient. For instance, Erdem et al. [22] develop a genetic algorithm to minimize the cost incurred due to disruptions of non-elective arrivals. Decisions that need to be made are whether or not to admit or defer the non-elective patient and how to adapt the elective schedule upon the arrival of the non-elective patient. They take the cost of postponing or preponing as well as the cost of turning down a patient into account. They deal with

one non-elective arrival per day at the beginning of the day. Another example is provided by Stuart and Kozan [84], who re-optimize the sequence of surgeries to incorporate the incoming non-electives by modeling a single-machine scheduling problem with due dates. They include sequence dependent processing times. Other approaches that are common for online rescheduling (schedule repair) are right shift scheduling, partial rescheduling and complete regeneration [98].

Inserting non-electives

A second option, next to reserving slack, is to incorporate the decision on when and how to schedule the non-electives into the sequencing or scheduling phase for the electives. In other words, the focus is on scheduling the electives in such a way that non-electives can be inserted in the schedule in the best way on the day of surgery (option 2 in Figure 1).

Building on the results of Wullink et al. [103], van Essen et al. [92] explore the option of break-in-moments for non-electives. The goal is to sequence the surgeries in their assigned OR such that the maximum interval between two consecutive BIMs is minimized. Several constructive and improvement heuristics are applied and tested. The non-elective waiting time can be reduced by spreading the BIMs as evenly as possible over the day. For an instance with sixteen ORs, the percentage of non-electives that have to wait longer than 30 minutes is smaller than 0.5%. Wullink et al. [103] also advocate the flexible approach thanks to the reduced waiting time for non-electives. However, in their case setting the worst results in the dedicated policy setting showed a waiting time of less than seven hours for all non-electives, which is still within the target window according to several classifications (see Table 4 earlier).

Instead of inserting the non-electives directly into the schedule, another approach is to assign 'buffers' in which the non-electives can be served if needed. This is different from reserving a certain amount of slack (e.g., 10% of capacity) since the buffers are spread out over the ORs and over time (see Figure 1, option 2) and can be variable in size. These buffers protect against unforeseen non-electives, but can also protect against duration variation. In the search for the best spots in the schedule to insert breaks, the research of Klassen and Rohleder [39,40] can provide insights. They study outpatient appointment scheduling and conclude that leaving open slots for urgent patients at the end of the day improves both the percentage of the urgent customers served and the server idle time while open slots at the beginning of the day decrease customer waiting time, but also decrease the percentage of urgent customers served. Moreover, they argue that it is best to evenly spread the open slots over the day, which is similar to the results on BIM. These findings could be tested in an inpatient surgery scheduling environment in future research.

Pham and Klinkert [72] adapt the multi-mode blocking job shop problem to the healthcare context and develop a MIP formulation. They model the problem of inserting non-electives as the job insertion problem. They insert the non-elective patients such that the resource assignments and the patient (job) sequence remain the same. They include not only the OR as a resource, but also nurses, surgeons, anesthetist and downstream capacity.

Several hospitals also use heuristics to 'plan' or insert the non-electives. For instance, Azari-Rad et al. [5] report that non-electives who need surgery within two hours of their arrival are assigned to the first available OR. If the arriving non-elective needs to be served between two and eight hours from the arrival time, the patient is assigned to the end of the day (until 11 PM) or is served as the first case of the next morning.

Furthermore, the scheduling of add-on cases has been researched by some authors (e.g., [19,105]). Dexter et al. [19] evaluate ten scheduling algorithms, which are online and offline variants of the best fit and worst-fit algorithms, to schedule add-on cases in the 'remaining' OR time. Hence, the algorithms are suited for solving the variable-sized bin packing problem with bounded space. Their result is applicable to ORs that schedule one or zero add-on cases per OR day. The days with open OR time, on which add-ons can be scheduled, are derived from empirical data. In the case hospitals (one covering 22 inpatient ORs and one having six outpatient ORs) there was on average more than an hour remaining in each OR per day. In about 45% of the ORs there was no remaining time. Scheduling the add-on cases increased the utilization in both hospitals from around 84% to about 93%.

6. Dedicated policy

In the dedicated policy, a subset of ORs is dedicated to serve a specific group of patients. This policy fits into the idea of reducing flow variability. Long waiting times are often caused by a flow problem and not by a resource problem [31] and separating electives and non-electives is one way to reduce the flow variability. The purpose of a DOR is often to improve access to care for both elective and non-elective patients and to reduce rescheduling and cancellation actions. This policy is in line with the recent guidelines published by the Royal College of Surgeons to separate both patients flows [86] and with the idea of reducing system-wide variability by separating the flows [31,57].

From the perspective of the elective patients, capacity is reduced in both the flexible and the dedicated scenario. In the former, emergency disruptions indirectly reduce capacity [23] while in the

latter the capacity is directly reduced. The main question is which reduction is the best option. Dhupar et al. [21] show for instance that delayed OR availability for urgent surgeries, which is the highest during regular operating hours, significantly increases the total hospital costs. They provide a retrospective study of five years of data for the appendectomy procedure.

When dedicating capacity, it is important to decide for which patient categories the capacity will be reserved. Although the goal is mostly to reserve capacity for non-electives (e.g.,[23,32,55,89]), the access to the DOR might be limited to trauma orthopedic patients (e.g., [11]), to urgent and semi-urgent patients (e.g., [9]), to add-on cases or to a combination of patient types (e.g., [75,79]).

A special case of DORs is when almost all ORs are dedicated to non-elective arrivals. This happens for instance in case of a disaster, where many rooms are cleared for incoming life-threatening cases.

6.1. Results

When using DORs, a clear trade-off between the reduction in flexibility and the increase in access time appears. In other words, it is the trade-off between a decrease in the waiting lists and the reduction of cancellations and overtime. In general, papers discussing the impact of DORs only report a very limited set of performance measures, which makes it difficult to properly assess the full range of the impact. The performance measures that are most used are the non-elective waiting time, the staff overtime, the OR utilization and the disruptions, as shown in Table 9. Note that comparing the results of the patient waiting time (for both categories) and the staff overtime for the dedicated policy with the results of the flexible policy (Table 8) are contradictory.

Firstly, the impact on the staff overtime is positive. However, many of the papers report mainly on a decrease in the number of non-electives served in overtime, but fail to report on the overall impact on overtime, which includes both the overtime of the electives and non-electives as well as the range of regular hours. This lack of information skews the interpretation of the results significantly.

Secondly, by separating the inherent variability from unscheduled emergency cases, the use of the elective ORs can be maximized. The utilization of the regular (elective) ORs is higher thanks to the reduction in disruptions. However, the utilization of the DORs ranges from 24% to 85% and depends on several factors like case mix, patient volume and scheduling policies.

Finally, splitting the flow of non-electives from the one of electives causes fewer unscheduled events in the elective schedule, which is consistent with the results from the flexible policy. This

means that fewer electives must be cancelled due to the arrival of non-electives and fewer rescheduling actions (to another OR) occur.

Table 9: Performance measures in a dedicated policy

Performance measure	Increase/decrease	Reference
Non-elective waiting time	↓/↑	[32,69,75,79][c] /[23]
Staff overtime	↓	[9,23,32,71,74,75,79]
Night time surgery	↓	[7,32,55,59,75,79]
Utilization of elective ORs	↑	[23,79]
Utilization of non-elective ORs	[24%, 37%, 42%, 53%, 85%, 88%] [b]	[7][75][32][79][55][9][d]
Disruptions[a] (electives)	↓	[9,32,55,71,79]
Elective patient waiting time	↑	[23,71]
Throughput	↑	[71,75,79]

[a] Disruption includes the number of disruptions, the number of rescheduled patients and the number of cancellations.
[b] [79] report a utilization of 33.5% according to the model and 60% observed after implementation of the DOR. [23] report a utilization between 24% and 75%.
[c] [32] report only the waiting time for priority three patients.
[d] References are in the order of the (increasing) utilization rates.

Other results that are mentioned are the decrease in length of stay, since the non-electives are served earlier in the day [32,74], a decrease in complications [9,32,74] and even morbidities [74].

Interestingly, some authors show that adapting the opening hours results in better system performance (e.g., [71]). As such, Smith et al. [79] propose to extend the opening hours of the OR for work-in cases by two hours (compared to the regular working hours) and the emergency ORs in [82] are available 24 hours a day, seven days a week. Additionally, Steins et al. [82] show that the opening hours played a key role in the reduced cancellations and the increased utilization.

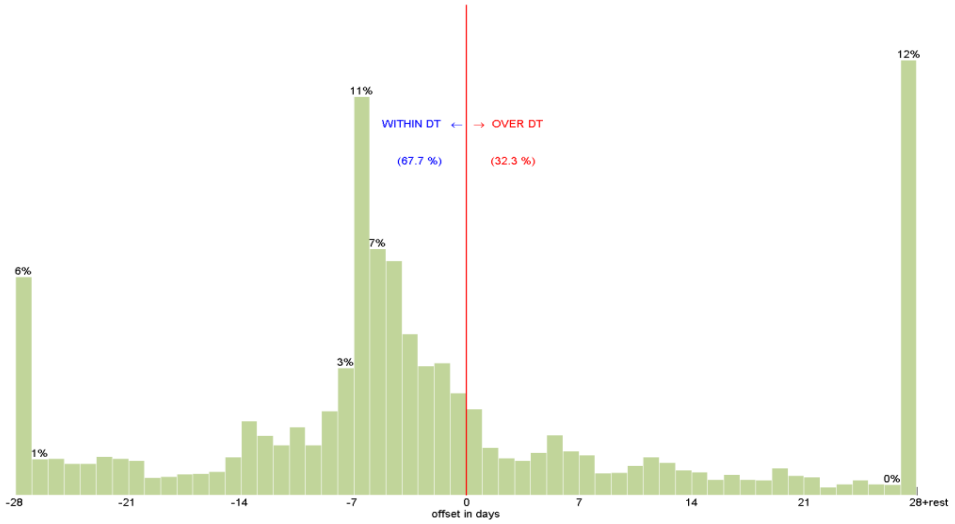
Concerning the waiting time, several remarks need to be addressed that contribute to the difficulties in comparing or assessing the results of the papers on dedicated capacity. First, the waiting time for both elective as well as non-elective patients is discussed in a surprisingly low number of papers. To assess the impact of introducing dedicated capacity, both performance measures need to be quantified. Similarly, the number of papers that clearly mention all elements of the trade-off between the utilization, the number of DORs and the waiting time for all patient categories is limited.

Secondly, a commonly used measure is the average waiting time over all patients. This measure is sensitive to outliers and to differences between the patient categories. An additional element that should accompany this measure is the distribution of the waiting time.

A third remark relates to the hospitals working with a classification system that prescribes a waiting time target for each category. In this case, the average waiting time does not provide the right information. After all, waiting for an average of twelve hours might be acceptable and even desired for some patients. Therefore, it might be more insightful to know whether or not a patient is served within his/her waiting time target (e.g. [69]), regardless of the total waiting time. Analogously to the second remark, unfortunately only a few of the papers using a target time window mention the distribution of the deviations from the target. This information would give a deeper understanding of the performance of the system with regards to waiting time and how well the hospital serves each patient group.

As an example from [76], Figure 2 shows the deviation in days from the due time target and provides a clear view on the percentage of patients that are served outside the due time interval. The authors also discuss the distribution of the offset per due time category. Sandbaek et al. [77] provide another example of the percentage of patients served within their service target per urgency category in a hybrid policy setting (discussed later), but do not show the distribution of the offset.

Figure 2: Offset from the due time target in days [76]



Finally, the waiting time for non-elective patients greatly depends on the exact implementation of the policy. For instance, if only one room is dedicated out of twenty ORs, the waiting time for non-electives will probably increase drastically (considering the patient mix in section 4.5). After all, they now only have a small part of the capacity, which they access in a highly variable way, resulting in a queue for the non-electives. However, dedicating more rooms in the same setting might have a positive effect on the waiting time of the non-electives.

A practical drawback of this policy is that if the emergency staff is sent away to deal with staff shortages in the elective ORs, the 'emergency' team might be incomplete upon arrival of a non-elective [103].

Finally, several papers discuss a setting applying the dedicated policy in the sense that the MSS includes one or more OR-days dedicated to non-elective patients (e.g., [2,10,43,78,82,91]). Unfortunately, these papers take this division as given. The same holds for the amount of dedicated capacity for the non-electives.

6.2. The handling of non-electives

The focus of projects and literature on DORs is on the identification, quantification and elimination of artificial variability in order optimally manage the flow of elective patients.

Smith et al. [79] separate the patient flows that cause the natural and artificial variation and test different room allocations for elective and non-elective patients. As such, they separate emergent and urgent cases, work-in cases (who need to be served within one week) and elective cases. Different scenarios (e.g., dedicating three compared to two rooms to all urgent and emergent surgeries) results in significantly different performance (e.g., lower average OR utilization). Subsequently, they allocate sufficient block time for 125% of each service's current demand in the elective rooms in order that, ceteris paribus, 80% of the allocated (elective) block time would be used. Next, they assign block time in order to smoothen the volumes throughout the week. In addition to the results reported in Table 9, they report a significant decrease in the day-to-day variability and an increase in net operating income. They report a use of five DORs after implementation.

On a smaller scale, Lovet and Katchburian [59] and Barlow et al. [7] show that the introduction of an afternoon emergency theater list, which is coordinated by a consultant anesthetist in [59], drastically improved the number of non-electives performed during daytime. However, Barlow et al. [7] report a utilization of the DOR of only 37% and only 36% of the non-elective operations were performed in the afternoon session while the majority was still performed in overtime, which makes the DOR a costly measure. They also show that this room was mainly used by urgent and scheduled cases (i.e., cases that need surgery within three weeks) and less by emergent cases.

Another option is the installation of a full-time DOR for add-on cases, where the non-electives are split up in different urgency categories, with a corresponding waiting time target (at most one day) (e.g., [32,69,75]). Using a combination of queueing theory and simulation, they determine the number

of DORs. In the model of Paul and MacDonald [69] patients can transition to a higher priority class if they experience excessive waiting. Heng and Wright [32] suggest to restrict the access of the add-on room for cardiologic and transplant cases because of their highly unpredictable nature and long surgery duration.

Heng and Wright [32] show that the lowest priority non-electives, priority 3 patients who need surgery within twelve hours, experience less waiting time, are served more within their target waiting levels and are operated on less in evening and night shifts. The DOR has little effect on access for priority 1 and 2 patients. Leppäniemi and Jousela [55] show similar results after the implementation of a classification system with three categories for non-electives and the introduction of a DOR for non-electives. The DOR leads to fewer elective cancellations and less overrun minutes in the elective rooms. Surprisingly, the number of delayed electives due to add-on cases remained similar after the introduction of an add-on room. Paul and MacDonald [69] show that the percentage of patients within their waiting limit does not depend on the patient mix between non-electives or on the surgery duration variations within the total volume. However, they assume that all surgeries follow the same duration distribution. Finally, Ryckman et al. [75] show that the schedule is more predictable by the introduction of an add-on room.

In the papers that focus on only a few departments (see Table 1), several authors discuss the option of dedicating an OR to orthopedic (trauma) patients. In the orthopedic department, the non-elective waiting time is more flexible since most patients can wait 24 hours before getting surgery. However, the orthopedic trauma cases are often waitlisted and served in overtime [9].

Bhattacharyya et al. [9] devote one OR to less urgent orthopedic cases. The room is controlled by orthopedics to schedule cases in a priority order determined by the day's attending staff surgeon. Persson and Persson [71] also devote one OR and introduce stand-by patients who can be called upon (on pre-defined dates that fit for the patient) when there is a free spot to go along with the DOR for emergencies. Similarly, Bhattacharyya et al. [9] mention that the DOR can be used for other cases if nothing is planned before 7 AM. If necessary, the emergency cases can also be planned in ORs that would be underutilized otherwise. The analysis in [9] only focusses on two common surgical cases and [71] consider a setting with two ORs.

Persson and Persson [71] model an optimization model (a MIP) and incorporate it in a simulation model. The optimization model consists of a bin packing model that minimizes the cost of surgery,

including a non-linear, non-decreasing cost related to the waiting time and costs with regards to cancellation and overtime.

Results show a significant shift to performing non-electives during daytime. This reduces complications since night time cases last longer due to the sometimes difficult surgeries with inexperienced staff on odd hours [9]. Surprisingly, Bhattacharyya et al. [9] report a utilization of 88% for the DOR, while the average for all services is 80%. Persson and Persson [71] also report a significant decrease in the number of surgery cancellations and overtime at the expense of an increase in waiting time for the electives.

Finally, Ferrand et al. [23] developed a simulation to compare the flexible and dedicated resource allocation policies. They assume that the elective schedule is fixed and elective patients arrive in batches according to fixed inter-arrival times. They collect the average and the maximum patient waiting times as well as the proportion of non-elective patients waiting longer than 30 minutes. The authors show that the reduced OR hours allocated to electives are more than compensated by the elimination of emergency disruptions in the dedicated policy. However, waiting time for non-electives increases and 30% of the non-electives waited for more than 30 minutes (target). The size of the effect is dependent on the variability in the processing times where higher variability results in lower waiting times for non-electives. They conclude that for high process variability the flexible policy results in better results for both patient categories, while in the base case the DOR improves waiting time for electives at the cost of waiting time for non-electives. Interestingly, increasing the volume of non-electives did not affect the divisions of ORs (5 out of 20 ORs for non-electives) in the dedicated policy.

While the previous articles focus more on capacity decisions, Dexter et al. [18] look at different sequencing heuristics for urgent surgeries, implicitly assuming they are served in the same OR, and show that the optimal sequence depends on the chosen performance measure. However, most articles on operational scheduling policies are designed for the flexible policy.

6.3. Dedicated team

Not just physical resources can be dedicated to a specific patient group, but also human resources. This aspect is often overlooked in the literature dealing with operational OR planning. In other words, staff is often assumed to be available when constructing the schedule. Without going into detail on the staffing decisions, several options are discerned in the reviewed literature.

Generally, an on-call team including surgeons, anesthesiologists and nurses is available to assist in case of emergency needs. However, the reviewed papers include a few specific suggestions regarding dedicated surgical staff. Bhattacharyya et al. [9] describe a system where one staff orthopedic surgeon is designated to cover each day of the week in the dedicated orthopedic trauma OR. Still, flexibility in switching surgeons for a case is desirable. Surgeons usually are on call the night before their designated trauma room days. Heng and Wright [32] propose a similar system. An on-call system is also proposed by Persson and Persson [71] for weekend shifts.

Another possibility is to have a special team that is experienced in the most common emergency procedures over all disciplines [41]. If no non-electives arrive, this team is used for handling semi-elective procedures or for replacing sick colleagues. Furthermore, Ryckman et al. [75] make a distinction between divisions that experience frequent urgent needs where a surgeon is assigned to the non-elective cases and divisions with fewer urgent patients where an on-call system is used.

Finally, van Oostrum et al. [96] determine an optimal OR team composition for operating during night-time with the two-fold goal of minimizing staffing costs and providing care within a target time interval. They show that modeling these intervals can reduce the required staff while still ensuring good quality of care.

6.4. How many ORs to dedicate

The answer to this question does not seem to be straightforward. The proportion of ORs that are dedicated varies widely and the same holds for the applied methods. Additionally, the patient categories to which the capacity is dedicated also differ.

Wullink et al. [103] consider one DOR out of a total of twelve staffed ORs for an average of five non-elective patients per day (mean case duration of 126 minutes). The other eleven ORs serve on average 32 electives (mean case duration of 142 minutes). On the contrary, Leppäniemi and Jousela reserve two additional DORs for non-electives for the same total amount of ORs. Another examples of diverse allocation mechanisms is provided by Tancrez et al. [89] and Zhang et al. [104] who reserve both only one room out of respectively six and twenty ORs. Similarly, for respectively a proportion of non-electives (see Table 6) of 50% ([55]) and 14% ([23]), one fourth of the capacity is allocated to non-electives.

Most authors use DES and queueing theory to find the amount of DORs. A steady state queueing analysis is used in [23] to determine the minimum number of required DORs. Also a multiserver Markov queueing model [32] and queueing formulas for the case of one OR [42] are used.

Ferrand et al. [23] test in a simulation model various divisions of the twenty ORs between the two patient groups and trade off elective waiting time with non-elective waiting time. They opt for five DORs. An analysis with the average number of overtime patients and the average and maximum overtime gives the same results. They also show that with an increasing number of DORs for non-electives (ranging from three to seven), the overtime first decreases and then increases from five DORs onwards. Besides, when dedicating three ORs, the percentage of patients outside the target waiting time is higher for non-electives than for electives. This switches from four ORs onwards. Ryckman et al. [75] split up the DOR decision for the night, weekend and day period and base their decision on the results for the offset to the waiting time, the utilization and the chance of having at least one room available for non-electives.

Paul and MacDonald [69] calculate the required number of ORs based on the total patient volume and the average surgery time in order to meet the 5% threshold for the service target time for all patients. They use a probit regression to determine the relation between the required number of ORs, and the total patient volume and the average surgery time. Next, they determine the relation between the proportion of patients of each priority exceeding their waiting time target and the optimal number of ORs, the average surgery time and the total patient volume.

Other authors [51,77] calculate the number of DORs by dividing the total expected capacity required for urgent surgery by the capacity per OR and take the smallest following integer. The 'last' OR is then filled up with electives for the remainder of its capacity.

The decisions on how many ORs to dedicate is closely related to the patient categories that are targets for dedicated capacity. Smith et al. [79] not only dedicate rooms (two) to urgent and emergent cases (including heart and lung transplants) and prior night emergencies, but also to work-in cases of predefined specialties and abdominal transplant (two rooms) and to electives (fifteen rooms). Ryckman et al. [75] test different settings and propose a similar division for twenty rooms. Steins et al. [82] dedicate two out of eighteen ORs to non-electives and dedicate some other ORs to a specific type of surgery like eye and oral surgery. Likewise, the outpatients and inpatients ORs are separated. Additionally, they separate the streams for weekdays, weekends and nights.

Finally, note that the practice of releasing OR block time can change the allocations of capacity on the short term. The capacity is released under predefined circumstances and the released capacity can be dedicated to a specific group of patients or can be released as flexible capacity. For instance, Bhattacharyya et al. [9] report that the DOR is freed up for other services if nothing is scheduled by 7 AM or if the cases end before 5 PM. Similarly, Persson and Persson [71] introduce stand-by patients, who are ready for having surgery in case of leftover capacity.

7. Hybrid policy

The hybrid policy consists of a mix of dedicated and flexible resources. As such, some rooms can be allocated to one patient group, but others can be used by multiple categories. This policy tries to obtain a better trade-off between flexibility and access time than the previous two policies.

In the hybrid policy different issues need to be addressed, which are all researched to a limited extent in OR planning and thus form an opportunity for future research. First, the number of dedicated and flexible resources must be decided. Next, the rules for accessing the flexible capacity need to be outlined, just like in the regular flexible policy. The difference is, however, that in the hybrid policy a part of the patients (a predefined group) is already served in the dedicated capacity, which might influence the operational rules for the flexible capacity. Another option is to originally dedicate capacity, but to allow patients to overflow under specific conditions. Decisions on when and how to overflow become necessary. The next paragraphs briefly discuss these options.

First, two recent examples of a hybrid policy for a large number of ORs (>20 ORs) are provided by Ferrand et al. [22] and Sandbaek et al. [68]. Ferrand et al. [22] examine different configurations of flexible and dedicated rooms and show that this policy outperforms the other two policies in terms of improved waiting time for both patient categories and lower overtime. They provide guidelines for the case where a limited amount of ORs is dedicated and for the case where a limited amount of ORs is made flexible. They also emphasize the importance of incorporating the prioritization among patients for obtaining the results. Interestingly, they report also on the 75^{th} and 95^{th} percentile of the performance measures. Sandbaek et al. [68] show that by a mix of introducing a new classification system and a hybrid OR policy, the overtime decreased, the utilization for all OR types increased and the waiting time for non-electives decreased. However, since they also implement a new classification and booking system, this effect might be caused by other factors than the dedicated policy. Also Sandbaek et al. [77] provide information on both the median waiting time and the proportion of patients within their target time. Interestingly, the results differ for the three non-elective categories.

The setting of Zonderland et al. [91] actually also follows a hybrid policy since they assume that urgent patients are served in a separate OR, while the semi-urgent patients need to be fitted into the elective schedule. Since the main focus of the paper was on the latter, it is discussed in the section on flexible policies.

Second, no uniform rules exist for the overflow of non-electives to elective capacity either. The highest priority category of non-electives might still be served in the OR that is assigned to the non-electives' specialty in the MSS, instead of in the DOR (e.g., [89]), while the other urgency categories are always served in the DOR. Exceptionally, Zhang et al. [89] look at all relevant performance measures and build a MIP to allocate OR capacity to different types of surgeries. Unfortunately, they do not provide a comparison with the scenario without DOR. In addition, long, complicated emergency cases, such as transplant or cardiovascular surgery, might not be suited for a DOR since they will greatly disturb the flow for the other non-electives, especially when only one DOR is available. Often a separate room is reserved for these surgeries or the surgeries are done in the regular time of the discipline. Stanciu et al. [72] provide an example of reserving capacity for different patient categories, which they define as a combination of a surgery type and a reimbursement level, in order to maximize the expected revenue. The expected revenue of a class depends on the (stochastic) surgery duration and the profit per unit OR time. The amount of time reserved for a class can be used by that class and by all classes with a higher priority. They calculate the reserved capacity based on insights from the revenue management literature.

Tancrez et al. [77,78] also research a setting with one DOR and five or six flexible rooms in which non-electives can still enter with priority if the DOR is occupied. They examine the impact of stochasticity in surgery durations, unexpected non-elective arrivals and blocking due to a full recovery unit. Interestingly, most papers on the dedicated policy report a decrease in the overtime while Tancrez et al. [78] report an increase. They argue that a decrease in capacity for the electives comes at the cost of either an increase in overtime or fewer planned operations within an acceptable overtime. The authors extend their first model [77] and include blocking in the OR because of a full recovery unit, together with an illustration of a real hospital case. They use a continuous Markov model. They quantify the disruptions of the schedule by the non-electives and show that the disruptions increase more than proportionally with the number of arriving non-electives. Dedicating an OR decreases the disruption rate, as well as the average non-elective waiting time and the probability of being served within 30 minutes.

Third, the DOR for non-electives might be used by electives, but clear guidelines on when and how to arrange this are lacking in the literature. One of the limited examples is provided by Bower and Mould [10], who install an orthopedic trauma session each day covering five hours out of the available seven hours. They suggest to schedule the remaining two hours for electives. The elective patients then accept a probability (15%) of being rescheduled in return for an earlier treatment. This results in a mix of dedicated capacity (five hours of trauma sessions) and flexible capacity (planned for electives, but still available for trauma cases if required). The authors show that a significant increase in elective throughput and utilization of the DOR can be achieved by this allocation.

8. Challenges and future directions

With regards to the problem of dealing with non-electives, the different policies are not yet fully explored. Although research has shown the impact of several policies or initiatives, the impact of different settings on the full spectrum of performance measures remains unclear. As a result, this is a possible area for future research. Furthermore, the literature on both the required slack as well as on inserting breaks is scarce. A key question that remains unanswered in the literature is how to determine what level of demand and which patient mix is required before a dedicated or flexible policy should be pursued. One could even think of a policy that pushes the described policies to an extreme, where resources are fully separated (physically). A recent example [16] shows the results for separating urgent and elective gynecology services.

Although the operating room and its organizational challenges have drawn considerable attention from researchers, there are still some challenges for future research. These challenges cover both the methodological perspective as well as the practical applicability.

From a methodological point of view, there is clearly a need for the development of appropriate stochastic methods. So far, stochasticity has been taken into account in several ways, from analytical methods (e.g., the newsvendor model) to approximations and sampling. Several papers use tailored heuristics to overcome the computational challenges.

From a modeling perspective, the master surgery scheduling problem (MSSP) and the surgery assignment problem (SAP) are well-known in the surgery scheduling literature. Although the precise problem focus might differ from paper to paper, it might be useful to incorporate the non-elective aspects into these problems as extensions and develop a common understanding of the problems and their extensions.

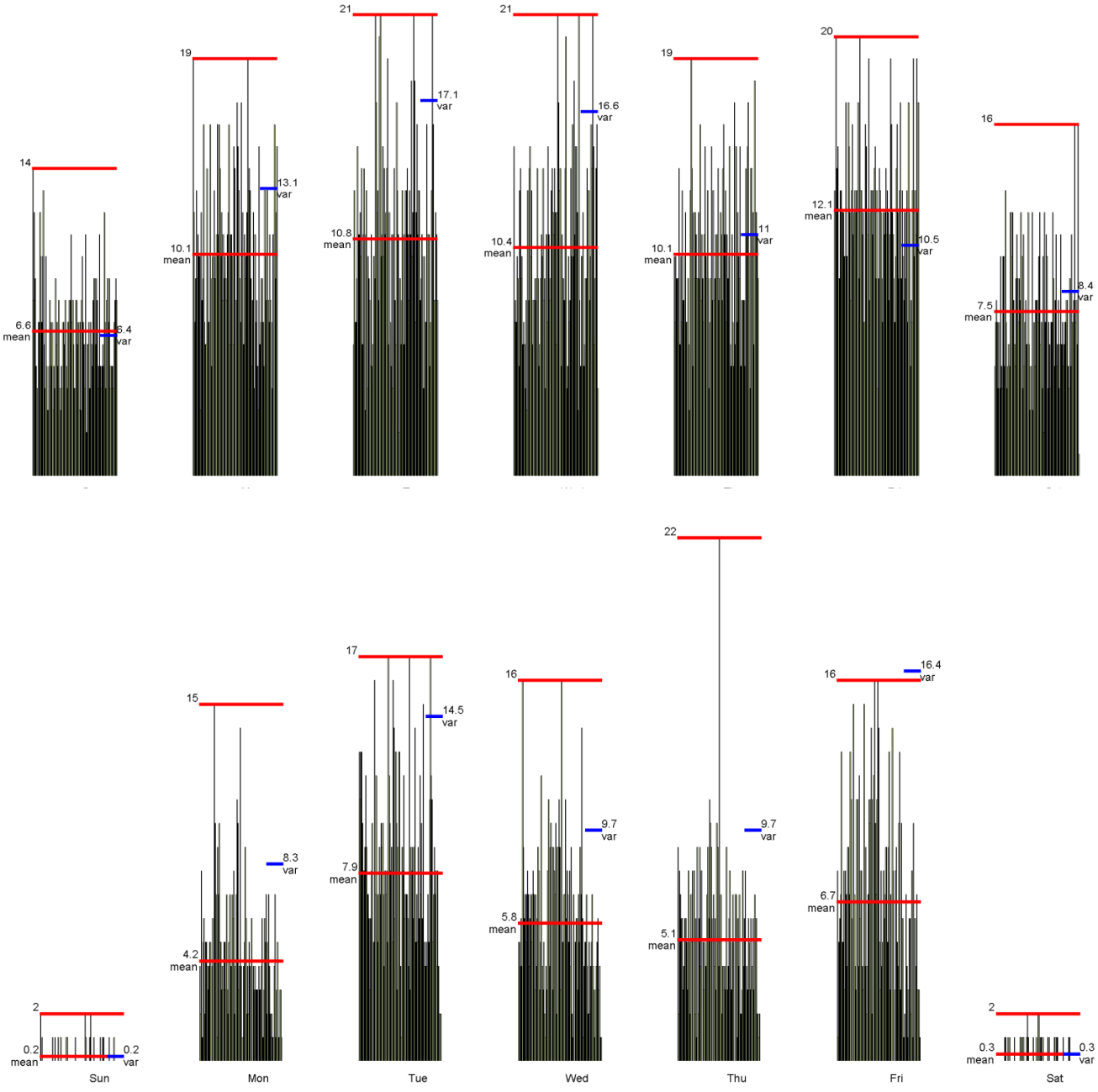
Generalizability is another issue in a lot of research on patient scheduling and capacity planning, as confirmed by the classifications in section 4. Researchers could develop a framework for the hospital setting in order to standardize the reporting on operating room planning topics. This would extent or complement the already existent frameworks on the problem setting (e.g., [14,99]). Additional information and testing on for instance the available OR time (OR block size), the relation between capacity and demand (e.g., an overcapacitated or a highly loaded system) and the scalability of the presented model is currently often lacking.

Unlike in the fields of for instance project scheduling, there are almost no test instances provided that cover realistic hospital settings, which could be used by researchers to align the research. Additionally, the disclosure of the dataset used is almost non-existent.

Making the research in such a way that it is practically applicable and academically sound has been a challenge for the operations researchers in the healthcare field. As shown in section 4.3, setting assumptions that cover a realistic setting remains a challenge. Furthermore, a thorough validation and/or motivation of the assumptions is often lacking, but highly relevant to judge the applicability. Although the assumptions on arrival and duration characteristics are crucial for interpreting the results, several papers fail to specifically mention them. Reporting all assumptions clearly should receive special attention in future research. New extensions such as introducing time-varying arrivals (which are common in the ED literature, but not in surgery scheduling) could also be considered more often in future research.

An example can show the importance of the point of verification and validation of the assumptions. Figure 3 shows the patient arrivals per weekday of a large university hospital with 22 ORs [76]. For a discipline like Urology, the variance clearly deviates from the mean, questioning whether the assumption of Poisson arrivals can be motivated. For the non-electives, shown at the top of Figure 3, the situation looks slightly better although the Poisson assumption does not hold for each weekday.

Figure 3: The number of arrivals for a given weekday per week (104 weeks for 2012-2013) for non-electives (top graph) and Urology (bottom graph) as well as the corresponding mean, maximum and variance [76]



Looking for more comprehensive performance measures that also take the differences between the patient categories into account is an area for future research. For instance, Tanaka et al. [87] develop measures that are adjusted for the hospital size and manpower to enable multi-institutional

comparisons. Additionally, Webster et al. [102] question whether policies focusing on the waiting time lead to patient-centered care. They argue that the waiting time does not describe the full experience of the patient and thus only covers one part of the service quality. As in inventory management, service levels try to capture the performance of the inventory policy, a comprehensive measure would contribute to tackling some of the challenges mentioned.

In the medical literature, patient safety, service quality, patient satisfaction and hospital costs (e.g., length of stay) often appear as performance measures. Moreover, from the perspective of the surgeons, financial pressure and clinical freedom influence managed care [85]. Performance measures such as mortality rate (e.g., due to late admission), morbidity, medical scores (e.g., injury severity score [61]), the number of complications during nights (related to the number of patients that can be scheduled within slot hours), the number of readmissions and the number of rescheduling actions are commonly used, while they appear less in the operations research literature. Also measures such as the quality adjusted life years (QUALY) are used to incorporate the quality aspect of medical care. For instance, Keren and Pliskin [38] look at the ideal (in terms of QUALY) timing for a patient to receive surgery. Finally, risk measures, which are popular in other areas, are not widely present in the discussed literature.

In general, the literature on how to adapt the elective schedule upon the arrival of non-electives is scarce and constitutes a great area for future research. One example is provided by Van Essen et al. [93], who provide a decision support system for the rescheduling problem in the OR. However, similar concepts can be found in other domains. For instance, in proactive-reactive project scheduling, a baseline schedule is created and a reactive policy is applied when during project execution disruptions such as longer activity durations or the unavailability of resources occur (e.g., [44,45,100,101]). Moreover, Artigues et al. [3] developed insertion techniques for the static and dynamic resource-constrained project scheduling problem. In the ED context, Lee [53] provides an analytic decisions framework that combines simulation optimization, machine learning and predictive analytics. Herroelen and Leus [33] provide an overview of reactive mechanisms in project scheduling.

Finally, in a broader scope, several other challenges lay ahead. The reviewed research focuses on reducing the negative effects of variability, which is often drawn from historical data and assumed to be input for the solution approach. However, developing or researching methods to reduce the variability itself (e.g., of late arrivals) must continue to gain equal attention. Initiatives focusing on a redesigned OR flow for emergencies can contribute to this.

Increasing the efficiency in handling the non-electives by decreasing the turnaround (e.g., parallel processing, specialized flow for emergencies), ensuring timely access for support services and minimizing the length of stay of elective patients can be explored further.

As shown in this paper, prioritization and categorization have been the subjects of several studies. However, a comprehensive overview of the impacts on scheduling performance measures of different categorizations could add to the current literature. Moreover, working towards a more standardized designation of the different patient categories can help both researchers and practitioners to increase the generalizability of the results.

Furthermore, Paul and MacDonald [69] also present algorithms that estimate the appropriate pricing for the surgeries, differentiated by priority level and given the patient demand and the resources reserved to meet this demand. The price is set such that a patient who waits longer due to a higher priority patient is charged less. Although setting pricing mechanisms is not feasible for each public financing framework, it might provide further possibilities for controlling the non-elective patient population.

Finally, the literature widely discusses the relation of OR planning with downstream resources such as the ICU and also the bed planning problem gained attention in the last decades. Looking at the source of non-electives, the emergency department plays an important role. Research on the ED is elaborate. However, a surprisingly low number of papers discusses the relation between OR planning and ED non-elective arrivals.

9. Conclusions

This paper reviews the trade-offs in the OR planning for elective and non-elective patients at both the tactical and the operational level. These trade-offs are caused by to different sources of variability. In general, three policies are pursued to handle the trade-offs: the flexible, dedicated and hybrid policy. They differ in the amount of capacity that is dedicated to a certain patient category. The majority of the research focuses on the flexible and the dedicated policy. The hybrid policy only recently received more attention and is a possible area for future research.

Most hospitals want to ensure acceptable access time while still keeping the resource utilization high. What acceptable means, depends on the patient classification used in the hospital. For the hospitals including target waiting times a standardized classification is lacking, which reduces the transferability of the results. Unfortunately, the setting of the reviewed papers differs a lot, which

renders a comparison difficult. For instance, not all papers look at all departments in the hospital, but focus on one or a few department(s). This raises the question whether different departments might be better suited for different policies. In addition, the modeling assumptions vary widely.

The main trade-offs include overtime, utilization, schedule disruptions and waiting time for elective and non-elective patients. Pursuing a more flexible policy should lead to fewer schedule disruptions and higher overall utilization. However, the effect on overtime and waiting time for both patient categories is still unclear. Although the performance of all drivers is required to get a complete view on the performance of the policy, often the analysis is limited to only a few of them. In future research, looking for new and more comprehensive performance measures might be useful. Additionally, the relation between the policies and the patient mix, the duration characteristics and the available number of ORs can be further explored.

Going back to the original need of finding policies for handling the trade-offs in the OR, the review shows that this need is not yet fully satisfied. The dedicated and the flexible policy proved to be efficient for certain case hospitals. However, the impact of all the drivers (e.g., waiting time, utilization, overtime, cancellations) on the choice of the policies is researched to a limited extent. In addition, the benefits of hybrid policies are not yet clear. Next to the hybrid policy, the impact of inserting breaks in the schedule forms an area for future research. In general, a clearer view on how to divide capacity and which elements drive such decisions is required. New (stochastic) algorithms or methods could be developed for this purpose in future research.

Acknowledgments

We acknowledge the support given to this project by the Research Foundation – Flanders (FWO-Vlaanderen) as Aspirant (Carla Van Riet).

10. References

[1] Adan I., Bekkers J., Dellaert N., Jeunet J., Vissers J., Improving operational effectiveness of tactical master plans for emergency and elective patients under stochastic demand and capacitated resources. Eur J Oper Res 213(1) (2011) 290–308.

[2] Agnetis A., Coppi A., Corsini M., Dellino G., Meloni C., Pranzo M., Long term evaluation of operating theater planning policies. Oper Res Heal Care 1(4) (2012) 95–104.

[3] Artigues C., Michelon P., Reusser S., Insertion techniques for static and dynamic resource-constrained project scheduling. Eur J Oper Res 149(2) (2003) 249–67.

[4] Augusto V., Xie X., Grimaud F., A framework for the modeling and simulation of health care systems. In: Proceedings of the 3rd Annual IEEE Conference on Automation Science and Engineering. IEEE, 2007:231–6.

[5] Azari-Rad S., Yontef A.L., Aleman D.M., Urbach D.R., Reducing elective general surgery cancellations at a Canadian hospital. Can J Surg 56(2) (2013) 113–8.

[6] Barkaoui K., Dechambre P., Hachicha R., Verification and optimisation of an operating room workflow. In: Proceedings of the 35th Hawaii International Conference on System Sciences. 2002:1–10.

[7] Barlow A.P., Wilkinson D.A., Wordsworth M., Eyre-Brook I.A., An emergency daytime theatre list: Utilisation and impact on clinical practice. Ann R Coll Surg Engl 75(6) (1993) 441–4.

[8] Beliën J., Demeulemeester E., Building cyclic master surgery schedules with leveled resulting bed occupancy. Eur J Oper Res 176(2) (2007) 1185–204.

[9] Bhattacharyya T., Vrahas M.S., Morrison S.M., Kim E., Wiklund R. a, Smith R.M., Rubash H.E., The value of the dedicated orthopaedic trauma operating room. J Trauma 60(6) (2006) 1336–41.

[10] Blake J.T., Donald J., Mount Sinai hospital uses integer programming to allocate operating room time. Interfaces (Providence) 32(2) (2002) 63–73.

[11] Bowers J., Mould G., Managing uncertainty in orthopaedic trauma theatres. Eur J Oper Res 154(3) (2004) 599–608.

[12] Brailsford S., Overcoming the barriers to implementation of operations research simulation models in healthcare. Clin Investig Med 28(6) (2005) 312–5.

[13] Brailsford S.C., Harper P.R., Patel B., Pitt M., An analysis of the academic literature on simulation and modelling in health care. J Simul 3(3) (2009) 130–40.

[14] Cardoen B., Demeulemeester E., Beliën J., Operating room planning and scheduling: A literature review. Eur J Oper Res 201(3) (2010) 921–32.

[15] Cardoen B., Demeulemeester E., On the use of planning models in the operating theatre: Results of a survey in Flanders. Int J Health Plann Manage 25(4) (2010) 400–14.

[16] Choo T., Deb S., Wilkins J., Atiomo W., Evaluating the impact of the reconfiguration of gynaecology services at a University Hospital NHS trust in the United Kingdom. BMC Health Serv Res 14(1) (2014) 428.

[17] Dexter F., Macario A., Lubarsky D.A., Burns D.D., Statistical method to evaluate management strategies to decrease variability in operating room utilization. Anesthesiology 91(1) (1999) 262–74.

[18] Dexter F., Macario A., Traub R.D., Optimal sequencing of urgent surgical cases. J Clin Monit Comput 15(3-4) (1999) 153–62.

[19] Dexter F., Macario A., Traub R.D., Which algorithm for scheduling add-on elective cases maximizes operating room utilization? Anesthesiology 91(5) (1999) 1491–500.

[20] Dexter F., Shi P., Epstein R.H., Descriptive study of case scheduling and cancellations within 1 week of the day of surgery. Anesth Analg 115(5) (2012) 1188–95.

[21] Dhupar R., Evankovich J., Klune J.R., Vargas L.G., Hughes S.J., Delayed operating room availability significantly impacts the total hospital costs of an urgent surgical procedure. Surgery 150(2) (2011) 299–305.

[22] Erdem E., Qu X., Shi J., Rescheduling of elective patients upon the arrival of emergency patients. Decis Support Syst 54(1) (2012) 551–63.

[23] Ferrand Y.B., Magazine M.J., Rao U.S., Comparing two operating-room-allocation policies for elective and emergency surgeries. In: Johansson B., Jain S., Montoya-Torres J., Hugan J., Yüceson E., eds. Proceedings of the 2010 Winter Simulation Conference. IEEE, 2010:2364–74.

[24] Ferrand Y.B., Magazine M.J., Rao U.S., Partially flexible operating rooms for elective and emergency surgeries. Decis Sci 45(5) (2014) 819–47.

[25] Ferrand Y.B., Magazine M.J., Rao U.S., Managing operating room efficiency and responsiveness for emergency and elective surgeries: A literature survey. IIE Trans Healthc Syst Eng 4(1) (2014) 49–64.

[26] Gerchak Y., Gupta D., Henig M., Reservation planning for elective surgery under uncertain demand for emergency surgery. Manage Sci 42(3) (1996) 321–34.

[27] Green L., Queueing analysis in healthcare. In: Patient Flow: Reducing Delay in Healthcare Delivery. Springer US, 2006:281–307.

[28] Guerriero F., Guido R., Operational research in the management of the operating theatre: A survey. Health Care Manag Sci 14(1) (2011) 89–114.

[29] Gupta D., Queueing Models for Healthcare Operations. In: Denton B.T., ed. Handbook of Healthcare Operations Management: Methods and Applications. New York: Springer Science & Business Media, 2013:19–44.

[30] Hans E.W., Wullink G., Van Houdenhoven M., Kazemier G., Robust surgery loading. Eur J Oper Res 185(3) (2008) 1038–50.

[31] Haraden C., Resar R., Patient flow in hospitals: Understanding and controlling it better. Front Health Serv Manage 20(4) (2004) 3–15.

[32] Heng M., Wright J.G., Dedicated operating room for emergency surgery improves access and efficiency. Can J Surg 56(3) (2013) 167–74.

[33] Herroelen W., Leus R., Robust and reactive project scheduling: A review and classification of procedures. Int J Prod Res 42(8) (2004) 1599–620.

[34] Huschka T.R., Narr B.J., Denton B.T., Thompson A.C., Using simulation in the implementation of an outpatient procedure center. In: Mason S.J., Hill R.R., Mönch L., Rose O., Jefferson T., Fowler J.W., eds. Proceedings of the 2008 Winter Simulation Conference. IEEE, 2008:1547–52.

[35] Jahangirian M., Naseer A., Stergioulas L., Young T., Eldabi T., Brailsford S., Patel B., Harper P., Simulation in health-care: Lessons from other sectors. Oper Res 12(3) (2012) 45–55.

[36] Jun J.B., Jacobson S.H., Swisher J.R., Application of discrete-event simulation in health care clinics: A survey. J Oper Res Soc 50(2) (1999) 109–23.

[37] Katsaliaki K., Mustafee N., Applications of simulation within the healthcare context. J Oper Res Soc 62(8) (2011) 1431–51.

[38] Keren B., Pliskin J.S., Optimal timing of joint replacement using mathematical programming and stochastic programming models. Health Care Manag Sci 14(4) (2011) 361–9.

[39] Klassen K.J., Rohleder T.R., Scheduling outpatient appointments in a dynamic environment. J Oper Manag 14(2) (1996) 83–101.

[40] Klassen K.J., Rohleder T.R., Outpatient appointment scheduling with urgent clients in a dynamic, multi-period environment. Int J Serv Ind Manag 15(2) (2004) 167–86.

[41] Klimek M., Van Houdenhoven M., Ottens T., Operating theatres: Organization, costs and audit. In: Gullo A., ed. Anaesthesia, Pain, Intensive Care and Emergency APICE. Springer Milan, 2008:263–8.

[42] Knight V. a, Harper P.R., Modelling emergency medical services with phase-type distributions. Heal Syst 1(1) (2012) 58–68.

[43] Kuo P., Schroeder R., Mahaffey S., Bollinger R., Optimization of operating room allocation using linear programming techniques. J Am Coll Surg 197(6) (2003) 889–95.

[44] Lambrechts O., Demeulemeester E., Herroelen W., Proactive and reactive strategies for resource-constrained project scheduling with uncertain resource availabilities. J Sched 11(2) (2008) 121–36.

[45] Lambrechts O., Demeulemeester E., Herroelen W., Time slack-based techniques for robust project scheduling subject to resource uncertainty. Ann Oper Res 186(1) (2010) 443–64.

[46] Lamiri M., Dreo J., Xie X., Operating room planning with random surgery times. In: Proceedings of the 3rd Annual IEEE Conference on Automation Science and Engineering. IEEE, 2007:521–6.

[47] Lamiri M., Grimaud F., Xie X., Optimization methods for a stochastic surgery planning problem. Int J Prod Econ 120(2) (2009) 400–10.

[48] Lamiri M., Xie X., Dolgui A., Grimaud F., A stochastic model for operating room planning with elective and emergency demand for surgery. Eur J Oper Res 185(3) (2008) 1026–37.

[49] Lamiri M., Xie X., Zhang S., Column generation approach to operating theater planning with elective and emergency patients. IIE Trans 40(9) (2008) 838–52.

[50] Lankester B.J., Paterson M.P., Capon G., Belcher J., Delays in orthopaedic trauma treatment: Setting standards for the time interval between admission and operation. Ann R Coll Surg Engl 82(5) (2000) 322–6.

[51] Van der Lans M., Hans E.W., Hurink J.L., Wullink G., Van Houdenhoven M., Kazemier G., Anticipating urgent surgery in operating room departments. 2005. Working paper No.158.

[52] Lebowitz P., Why can't my procedures start on time? AORN J 77(3) (2003) 594–7.

[53] Lee E.K., Yuan F., Zhou R., Lahlou S., Post E., Wright M.D., Atallah H.Y., System analytics: Modeling and optimizing emergency department workflow. In: Hui Y., Lee E.K., eds. Healthcare Data Analytics. Wiley Series in Operations Research and Management Science. John Wiley & Sons, 2014

[54] Lehtonen J.-M., Torkki P., Peltokorpi A., Moilanen T., Increasing operating room productivity by duration categories and a newsvendor model. Int J Health Care Qual Assur 26(2) (2013) 80–92.

[55] Leppäniemi A., Jousela I., A traffic-light coding system to organize emergency surgery across surgical disciplines. Br J Surg 101(1) (2014) 134–40.

[56] Litvak E., Long M.C., Cooper A., McManus M.L., Emergency department diversion: Causes and solutions. Acad Emerg Med 8(11) (2001) 1107–10.

[57] Litvak E., Long M.C., Cost and quality under managed care: Irreconcilable differences. Am J Manag Care 6(3) (2000) 305–12.

[58] Litvak N., Rijsbergen M., Boucherie R., Van Houdenhoven M., Managing the overflow of intensive care patients. Eur J Oper Res 185(3) (2008) 998–1010.

[59] Lovett B.E., Katchburian M.V., Emergency surgery: Half a day does make a difference. Ann R Coll Surg Engl 81(1) (1999) 62–4.

[60] MacCormick A.D., Collecutt W.G., Parry B.R., Prioritizing patients for elective surgery: A systematic review. ANZ J Surg 73(8) (2003) 633–42.

[61] Matsushima K., Cook A., Tollack L., Shafi S., Frankel H., An acute care surgery model provides safe and timely care for both trauma and emergency general surgery patients. J Surg Res 166(2) (2011)

[62] McManus M.L., Long M.C., Cooper A., Litvak E., Queuing theory accurately models the need for critical care resources. Anesthesiology 100(5) (2004) 1271–6.

[63] Mielczarek B., Uzialko-Mydlikowska J., Application of computer simulation modeling in the health care sector: A survey. Simulation 88(2) (2010) 197–216.

[64] Min D., Yih Y., Scheduling elective surgery under uncertainty and downstream capacity constraints. Eur J Oper Res 206(3) (2010) 642–52.

[65] Min D., Yih Y., Managing a patient waiting list with time-dependent priority and adverse events. RAIRO-Operations Res 48 (2014) 53–74.

[66] Mullen P.M., Prioritising waiting lists: How and why? Eur J Oper Res 150(1) (2003) 32–45.

[67] O'Leary D.P., Beecher S., McLaughlin R., Emergency surgery pre-operative delays - Realities and economic impacts. Int J Surg 12(12) (2014) 1333–6.

[68] Patrick J., Puterman M., Reducing wait times through operations research: Optimizing the use of surge capacity. Healthc Policy 3(3) (2008) 75–88.

[69] Paul J., MacDonald L., Determination of number of dedicated OR's and supporting pricing mechanisms for emergent surgeries. J Oper Res Soc 64(6) (2012) 912–24.

[70] Peck J.S., Gaehde S. a, Nightingale D.J., Gelman D.Y., Huckins D.S., Lemons M.F., Dickson E.W., Benneyan J.C., Generalizability of a simple approach for predicting hospital admission from an emergency department. Acad Emerg Med 20(11) (2013) 1156–63.

[71] Persson M.J., Persson J.A., Analysing management policies for operating room planning using simulation. Health Care Manag Sci 13(2) (2010) 182–91.

[72] Pham D.-N., Klinkert A., Surgical case scheduling as a generalized job shop scheduling problem. Eur J Oper Res 185(3) (2008) 1011–25.

[73] Rachuba S., Werners B., A robust approach for scheduling in hospitals using multiple objectives. J Oper Res Soc 65(4) (2013) 546–56.

[74] Roberts T.T., Vanushkina M., Khasnavis S., Snyder J., Papaliodis D.N., Rosenbaum A.J., Uhl R.L., Roberts J.T., Bagchi K., Dedicated orthopaedic operating rooms: Beneficial to patients and providers alike. J Orthop Trauma 29(1) (2015) 18–23.

[75] Ryckman F., Adler E., Anneken A., Bedinghaus C., Clayton P.J., Hays K.R., Lee B., Morillo-delerme J.W., Schoettker P.J., Yelton P.A., Kotagal U.R., Cincinnati Children's Hospital Medical Center: Redesigning perioperative flow using operations management tools to improve access and safety. In: Litvak E., ed. Managing Patient Fow in Hospitals: Strategies and Solutions. Oakbrook Terrace, IL: Joint Commission International, 2009:97–111.

[76] Samudra M., First doctoral seminar. 2013 http://feb.kuleuven.be/healthcare/seminar_1/.

[77] Sandbaek B.E., Helgheim B.I., Larsen O.I., Fasting S., Impact of changed management policies on operating room efficiency. BMC Health Serv Res 14(1) (2014) 1–10.

[78] Santibáñez P., Begen M., Atkins D., Surgical block scheduling in a system of hospitals: An application to resource and wait list management in a British Columbia health authority. Health Care Manag Sci 10(3) (2007) 269–82.

[79] Smith C.D., Spackman T., Brommer K., Stewart M.W., Vizzini M., Frye J., Rupp W.C., Re-engineering the operating room using variability methodology to improve health care value. J Am Coll Surg 216(4) (2013) 559–68.

[80] Sobolev B.G., Brown P.M., Zelt D., Fitzgerald M., Priority waiting lists: Is there a clinically ordered queue? J Eval Clin Pract 11(4) (2005) 408–10.

[81] Stanciu A., Vargas L.G., May J., A revenue management approach for managing operating room capacity. In: Johansson B., Jain S., Montoya-Torres J., Hugan J., Yücesan E., eds. Proceedings of the 2010 Winter Simulation Conference. IEEE, 2010:2444–54.

[82] Steins K., Persson F., Holmer M., Increasing utilization in a hospital operating department using simulation modeling. Simulation 86(8-9) (2010) 463–80.

[83] Strum D.P., May J., Vargas L.G., Modeling the uncertainty of surgical procedure times. Anesthesiology 92(4) (2000) 1160–7.

[84] Stuart K., Kozan E., Reactive scheduling model for the operating theatre. Flex Serv Manuf J 24(4) (2011) 400–21.

[85] Sturm R., The impact of practice setting and financial incentives on career satisfaction and perceived practice limitations among surgeons. Am J Surg 183(3) (2002) 222–5.

[86] Surgeons R.A.C. of, The case for the separation of elective and emergency surgery. 2011 http://www.surgeons.org/media/307115/sbm_2011-05-24_separating_elective_and_emergency_surgery.pdf.

[87] Tanaka M., Lee J., Ikai H., Imanaka Y., Development of efficiency indicators of operating room management for multi-institutional comparisons. J Eval Clin Pract 19(2) (2013) 335–41.

[88] Tancrez J.-S., Roland B., Cordier J.-P., Riane F., How stochasticity and emergencies disrupt the surgical schedule. In: McClean S., Millard P., El-darzi E., Nugent C.D., eds. Intelligent Patient Management. Springer Berlin Heidelberg, 2009:221–39.

[89] Tancrez J.-S., Roland B., Cordier J.-P., Riane F., Assessing the impact of stochasticity for operating theater sizing. Decis Support Syst 55(2) (2013) 616–28.

[90] Testi A., Tanfani E., Valente R., Ansaldo G.L., Torre G.C., Prioritizing surgical waiting lists. J Eval Clin Pract 14(1) (2008) 59–64.

[91] Testi A., Tànfani E., Tactical and operational decisions for operating room planning: Efficiency and welfare implications. Health Care Manag Sci 12(4) (2009) 363–73.

[92] van Essen J.T., Hans E.W., Hurink J.L., Oversberg a., Minimizing the waiting time for emergency surgery. Oper Res Heal Care 1(2-3) (2012) 34–44.

[93] van Essen J.T., Hurink J.L., Hartholt W., van den Akker B.J., Decision support system for the operating room rescheduling problem. Health Care Manag Sci 15(4) (2012) 355–72.

[94] Van Houdenhoven M., Hans E.W., Klein J., Wullink G., Kazemier G., A norm utilisation for scarce hospital resources: Evidence from operating rooms in a Dutch university hospital. J Med Syst 31(4) (2007) 231–6.

[95] Van Houdenhoven M., van Oostrum J.M., Hans E.W., Wullink G., Kazemier G., Improving operating room efficiency by applying bin-packing and portfolio techniques to surgical case scheduling. Anesth Analg 105(3) (2007) 707–14.

[96] van Oostrum J.M., Van Houdenhoven M., Vrielink M.M.J., Klein J., Hans E.W., Klimek M., Wullink G., Steyerberg E.W., Kazemier G., A simulation model for determining the optimal size of emergency teams on call in the operating room at night. Anesth Analg 107(5) (2008) 1655–62.

[97] Vansteenkiste N., Lamote C., Vandersmissen J., Luysmans P., Monnens P., De Voldere G., Kips J., Rademakers F.E., Reallocation of operating room capacity using the due-time model. Med Care 50(9) (2012) 779–84.

[98] Vieira G.E., Herrmann J.W., Lin E., Rescheduling manufacturing systems: A framework of strategies, policies and methods. J Sched 6 (2003) 39–62.

[99] Vissers J., Bertrand J., De Vries G., A framework for production control in health care organizations. Prod Plan Control 12(6) (2001) 591–604.

[100] Van de Vonder S., Demeulemeester E., Herroelen W., Leus R., The use of buffers in project management: The trade-off between stability and makespan. Int J Prod Econ 97(2) (2005) 227–40.

[101] Van de Vonder S., Demeulemeester E., Herroelen W., A classification of predictive-reactive project scheduling procedures. J Sched 10(3) (2007) 195–207.

[102] Webster F., Perruccio A. V, Jenkinson R., Jaglal S., Schemitsch E., Waddell J.P., Bremner S., Mobilio M.H., Venkataramanan V., Davis A.M., Where is the patient in models of patient-centred care: a grounded theory study of total joint replacement patients. BMC Health Serv Res 13(1) (2013) 531.

[103] Wullink G., Van Houdenhoven M., Hans E.W., van Oostrum J.M., van der Lans M., Kazemier G., Closing emergency operating rooms improves efficiency. J Med Syst 31(6) (2007) 543–6.

[104] Zhang B., Murali P., Dessouky M.M., Belson D., A mixed integer programming approach for allocating operating room capacity. J Oper Res Soc 60(5) (2009) 663–73.

[105] Zhou J., Dexter F., Method to assist in the scheduling of add-on surgical cases: Upper prediction bounds for surgical case durations based on the log-normal distribution. Anesthesiology 89(5) (1998) 1228–32.

[106] Zonderland M., Boucherie R., Litvak N., Vleggeert-Lankamp C., Planning and scheduling of semi-urgent surgeries. Health Care Manag Sci 13(3) (2010) 256–67.